IJATE

Volume 4 Issue 2 2017



# International Journal of Assessment Tools in Education

International Journal of
Assessment Tools in Education

http://ijate.net/

e-ISSN: 2148-7456

# International Journal of Assessment Tools in Education (IJATE)

Publisher : İzzet KARA
Frequency : 2 issues per year
Online ISSN : 2148-7456
Website : http://www.ijate.net/index.php/ijate
Address : Pamukkale University, Education Faculty,
Department of Mathematics and Science Education
20070, Denizli, Turkey

*Phone:* +90 258 296 1036
*E-Mail:* ijate.editor@gmail.com

IJATE will be published biannual (one volume per year, two issues per year -January and July). IJATE welcomes the submission of manuscripts that meets the general criteria of significance and scientific excellence.

There is no submission or publication process charges for articles in IJATE.

***ISSN: 2148-7456***

**IJATE is indexed in:**

• DOAJ,
• Index Copernicus International,
• Google Scholar,
• Türk Egitim İndeksi,
• Open Access Journals,
• Akademik Dizin,
• Academic Keys,
• CiteFactor (ASJ),
• SIS (Scientific Index Service) Database,
• SCIPIO (Scientific Publishing & Information Online),
• MIAR 2015 (Information Matrix for Analysis of the Journals),
• I2OR Indexing Services,
• JournalTOCs,
• Sosyal Bilimler Atıf Dizini (SOBIAD),
• International Innovative Journal Impact Factor ( IIJIF)

- *International Journal of Assessment Tools in Education* (IJATE) is a peer-reviewed on-line journal.
- Author(s) are responsible from the copyrights of the figures, pictures and visuals, and the content of the manuscripts, accuracy of the references and quotations and proposed ideas.

ISSN: 2148-7456

**Volume 4, Issue 2, (2017)**

**Table of Contents**

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 96-114*

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

*"Research Article"*

# Developing a Proof-of-Concept Selection Test for Entry into Primary Teacher Education Programs

**Robert M. Klassen,[1] Tracy L. Durksen,[2] Lisa E. Kim,[1] Fiona Patterson,[3,4] Emma Rowett,[3] Jane Warwick,[4] Paul Warwick,[4] and Mary-Anne Wolpert[4]**

[1]Department of Education, University of York, York YO10, 5DD, United Kingdom
[2]University of New South Wales, Australia
[3]Work Psychology Group, UK
[4]University of Cambridge, UK

***Abstract***

The purpose of this article is to report on the development of a proof-of-concept situational judgment test (SJT) to assist in the selection of candidates for primary teacher education (ITE) programs. Nine development steps involving practising teachers, teacher educators, and applicants to ITE programs were carried out to establish target attributes and to develop content for the test. The results from administering the test to 124 primary ITE candidates showed a near-normal distribution, high levels of reliability, and significant positive correlations with a range of concurrently administered interview scores. We conclude with a description of the necessary next steps needed to implement evidence-supported teacher education selection processes in a range of international settings.

***Article Info***

***Received***
*26 September 2016*

***Revised***
*19 November 2016*

***Accepted***
*23 November 2016*

***Key words***
teacher selection;
initial teacher education;
situational judgment tests;
teacher effectiveness;
recruitment;
teacher characteristics

## 1. INTRODUCTION

Identifying and selecting the most promising prospective teachers has been a continuing challenge in educational research and practice for nearly 100 years (e.g., Knight, 1922; Staiger & Kane, 2015). Any selection process is built on an evaluation of data to make predictions about future effectiveness. Selecting candidates for initial teacher education (ITE) programs presents selectors with questions about the kinds of data to evaluate: Which characteristics of candidates should be evaluated? How can these characteristics be evaluated in a way that is reliable, valid, and fair? Are these characteristics associated with success in teacher education and teaching practice? The conventional selection approach for ITE programs is to ask candidates for some combination of academic transcripts, personal statements, letters of reference, and to participate in individual interviews. However, there is little evidence supporting the use of many conventional ITE selection procedures (Casey & Childs, 2011),

[1] Corresponding Author Phone: +44 07914 701260 Email: robert.klassen@york.ac.uk

2148-7456 /© 2017 DOI: 10.21449/ijate.275772

and furthermore, some selection methods-including interviews and letters of reference-may be unreliable and systematically biased against certain groups of candidates (McDaniel, Whetzel, Schmidt, & Maurer, 1994). In this proof-of-concept study, we report the development and initial evaluation of an innovative selection tool for use in selecting candidates for primary ITE programs.

### 1.1. The case for improving selection procedures into initial teacher education

High-performing education systems tend to place importance on developing effective ITE selection processes (Barber & Mourshed, 2007; Sahlberg, 2014; Sclafani, 2015), with selection methods that include evaluation of candidates' academic and non-academic attributes[1]. Researchers and policy-makers in a range of settings have called for improvements in ITE selection in efforts to improve teacher quality (Heinz, 2013; Thomson et al., 2011; UK House of Commons, 2012). In any jurisdiction, selection is necessary for three reasons: a) to make decisions about 'selecting in' when the number of applicants outweighs the number of available places, b) to make decisions about 'selecting out' in order to identify those candidates who may be unsuitable, and c) to provide a profile of candidates' strengths and weaknesses for future development. At the foundation of selection research is the belief that individuals vary in personal attributes and experiences, and that these individual differences are related to future behaviors in training and professional contexts.

Although almost all novice teachers become more effective with experience and professional training (Hanushek & Rivkin, 2011), their effectiveness relative to their peers remains quite stable over time (Atteberry, Loeb, & Wyckoff, 2015). That is, novice teachers' relative effectiveness is heterogeneous and is predictive of their future relative effectiveness, especially for those who initially display the highest and lowest levels of relative effectiveness (Atteberry et al.). Furthermore, although many candidates entering ITE programs will show growth in non-academic attributes (e.g., professional commitment and motivation) during the duration of their program, some candidates will show persistently low levels of professional commitment and motivation (e.g., Klassen & Durksen, 2014; Watt, Richardson, & Wilkins, 2014). Watt et al. (2014) traced the professional commitment and motivation of students from the beginning to the end of their ITE programs, and found that a sizable group-28% of participants in their study-began the program with low levels of motivation for teaching and maintained that profile until the end of the program. Given the relative stability of teacher effectiveness and non-academic attributes, selection methods used by ITE programs should make the best possible predictions about the motivation and effectiveness trajectories of prospective teachers.

### 1.2. Current approaches for ITE selection

Uncovering the within-teacher factors that lead to teacher effectiveness is at the heart of the ITE selection process. Although attempts have been made to improve and systematise selection practices, there is a dearth of valid tools to help admissions committees make these important selection decisions in ITE programs (Mikitovics & Crehan, 2002). Selection into ITE programs typically involves evaluation of three factors: (1) academic attributes (such as

[1]The term 'academic' attributes (sometimes referred to as 'cognitive' attributes) refers to variables that reflect reasoning skills (such as the Scholastic Aptitude Test, SAT) or academic achievement (e.g., GPA or past performance in particular academic areas). The term 'non-academic attributes' (sometimes referred to as 'non-cognitive' attributes) refers to within-person variables, which might include beliefs, motives, personality traits, and dispositions (e.g., Patterson, Zibarras, & Ashworth, 2016).

subject area knowledge using evidence from university transcripts and sometimes through a written response to a journal article); (2) background experience (using evidence from personal statements and reference letters); and (3) *non-academic* attributes (such as personality, motives, and dispositions using evidence from interviews, personal statements, and occasionally, personality tests).

Figure 1 provides a model with examples of how these three factors are measured and how they are linked to performance for selection into ITE programs. Although teacher education programs vary in the kinds of assessments that they use for assessing candidates, we know very little about the reliability, validity, and perceived fairness of these procedures. What links disparate selection methods together is the common goal to identify candidates who show higher, rather than lower, levels of academic and non-academic attributes.

In the UK, a recent survey of 74 university-based (ITE) providers (Klassen & Dolan, 2015) found that all programs assessed academic attributes through evaluation of university academic transcripts, and that almost all assessed non-academic attributes through a combination of individual and group interviews (97%), and evaluation of behaviour in group activities (62%). In North America, specific selection methods for ITE programs vary widely, but selectors typically rely on some combination of candidates' previous academic achievement, individual and group interview performance, personal statements, letters of reference, and in some cases, government-mandated standardized tests (Casey & Childs, 2007). Selection into highly competitive Finnish ITE programs includes evaluation of academic attributes such as academic achievement, but also non-academic attributes including personality and interpersonal skills (Sahlberg, 2014). Similarly, selection into competitive Singaporean ITE programs includes an evaluation of academic attributes such as grades and national exams, but also evaluation of non-academic attributes including motivation, passion, values, and commitment to teaching (Sclafani, 2015). Almost all selection approaches have the same goal—to identify candidates with the highest potential for success during the program and in teaching practice—but there is little evidence for reliability, validity, and fairness of these selection methods internationally (Hobson, Ashby, McIntyre, & Malderez, 2010).

### 1.3. Situational judgment tests

In fields outside of education, there has been a keen interest in the use of situational judgment tests (SJTs) for employee selection, but also for selection into professional training programs, especially in medicine (e.g., Patterson, Zibarras, & Ashworth, 2016). SJTs are a measurement method designed to assess candidates' judgments of the benefits and costs of behaving in certain ways in response to challenging contextualised scenarios. In some ways, SJTs resemble a conventional face-to-face interview where a scenario might be presented orally to candidates with an open-ended response format (e.g., *Describe what you would do if*…). SJTs, however, differ from conventional interviews in that a larger sample of scenarios can be administered to applicants, the scoring key can be standardized, and the tests can be used to screen large numbers of applicants economically and efficiently. The format of SJTs can be in paper-and-pencil, computer-administered, or video-based. The development of SJT content is typically based on job analysis and through gathering 'critical incidents' from those already in the job (Patterson et al., 2016). Experienced professionals, or 'subject matter experts,' are used to generate response options (Lievens et al., 2008). Final scoring keys, which indicate more and less effective response options, are established through consensus with a panel of experts.

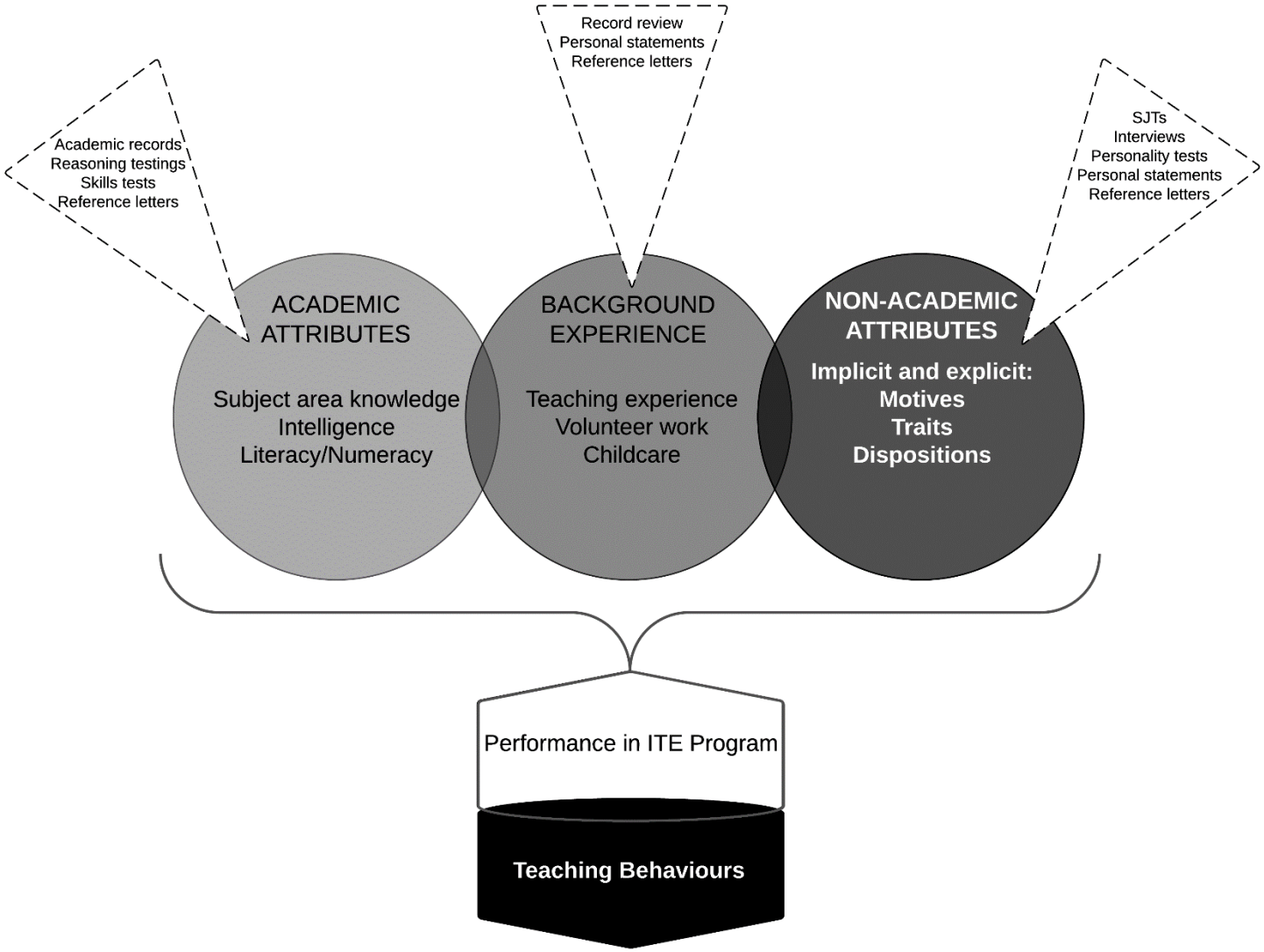



**Figure 1.** Model of relationship between academic attributes, background experience, and non-academic attributes in prediction of performance of ITE performance and teaching behaviors.

SJTs are designed to measure *implicit trait policies*; that is, the tendency individuals have to express traits in certain ways under particular contexts (Motowidlo & Beier, 2010). According to this theory-similarly conceptualised as tacit knowledge in Sternberg's theory of successful intelligence (e.g., Elliott, Stemler, Sternberg, Grigorenko, & Hoffman, 2011)-those who are more experienced in a particular job are more likely to implicitly understand optimal behavioral responses. However, novices with limited experience also have partial knowledge about effective response patterns, based on their implicit traits and understanding of the kinds of behaviors that are likely to be most appropriate in SJT scenarios (Motowidlo & Beier). In education, candidates for ITE programs have pre-existing beliefs about how to react to classroom challenges (e.g., how to manage classroom discipline issues), based on the procedural knowledge gained from their own life experiences, even when they do not have direct experience with teaching. These existing beliefs, or implicit trait policies, may change as candidates gain pedagogical knowledge and teaching experience, but remain as influences of teaching behaviors.

SJTs tend to display stronger face and content validity than conventional non-academic measures due to their close correspondence to the work-related situations that they describe (Whetzel & McDaniel, 2009).The interest in SJT methodologies is due to the promise of predictive validity (Patterson et al., 2016), with SJTs administered at admissions to medical school predicting job performance ($r$ = .22) nine years later (Lievens & Sackett, 2012). In a recent meta-analysis on SJT validities and reliabilities, Christian et al. (2010) found SJTs measuring interpersonal attributes had a mean validity coefficient of .25, those measuring conscientiousness had a mean coefficient of .24, and heterogeneous composite SJTs showed a mean validity of .28. A previous large-scale meta-analysis of SJT validity ($N$ = 24,756) using mostly concurrent validity studies showed a validity coefficient of .26 (McDaniel, Hartman, Whetzel, & Grubb, 2007).

Non-academic attributes can be measured using conventional, explicit measures of personality (e.g., 'How much is this statement like you?' *I am generally agreeable*) that are prone to socially desirable response patterns (Greenwald & Banaji, 1995; Johnson & Saboe, 2011). In contrast, SJTs can provide an indirect or implicit measure of what candidates view as appropriate ways of behaving in certain contexts (Motowidlo & Beier, 2010). Moreover, SJTs constructed in collaboration with expert practitioners are less susceptible to coaching effects and faking than many other kinds of selection tests because they are cognitively complex and are designed to measure implicit traits (Whetzel & McDaniel, 2009).

Researchers have also noted weaknesses in the research underpinning the development and use of SJTs for selection (e.g., Lievens, Peeters, & Schollaert, 2008). The vast majority of SJT validation studies have used a concurrent design with few studies establishing predictive validity (Campion, Ployhart, & MacKenzie, 2014). Although SJTs are often constructed to target particular attributes (e.g., professional integrity in medical selection; Patterson et al., 2016), their hypothesized factor structure is frequently not replicable in factor analysis (Lievens et al., 2008). In addition, internal consistency may be below conventional standards, and some SJTs have been shown to be prone to faking and coaching (Whetzel & McDaniel, 2009). SJTs are typically developed to reflect multiple dimensions, but because the content of individual items (scenarios) may reflect multiple dimensions, establishing the factor structure can be a challenge (Schmitt & Chan, 2006).

SJTs have been shown to predict performance in dentistry and medical training programs over and above cognitive measures (Lievens & Sackett, 2012; Patterson, et al, 2012). In the United States, SJTs were found to be a better predictor of lawyer effectiveness than the conventional tests used for selection into highly competitive law schools, and to be less prone

to inter-group bias (i.e., race, gender) than other measures (Shultz & Zedeck, 2012). Overall, SJTs have shown strong concurrent validity, some evidence of predictive validity (Lievens & Patterson, 2011), and a higher degree of fairness (i.e., less systematic bias) than other selection methods (Shultz & Zedeck, 2012).

***Current Study.*** SJTs are often designed deductively (top-down) to capture personality traits, but can also be designed to measure inductively-developed, contextualised non-academic attributes related to professional effectiveness. The current study describes the development and initial validation of a proof-of-concept SJT designed to be used for selection into primary level teacher education programs in the UK. Four research questions were posed:

(RQ1) Can a set of robust target attributes be established based on an inductive (bottom-up) approach?
(RQ2) Can an SJT developed for entry into primary ITE show acceptable psychometric properties?
(RQ3) Is the SJT a valid selection method (i.e., does the SJT show concurrent criterion-related validity with scores from the existing selection process)?
(RQ4) Do candidates view the SJT as fair and as a feasible selection method (i.e., does the test show face validity)?

## 2. METHOD AND RESULTS

The ITE selection SJT was designed to assess non-academic attributes required for success as a novice teacher in UK primary schools. We followed best-practice approaches to SJT development from the organizational psychology literature (Campion et al., 2014), and in particular, the approach used by Patterson et al., 2015 as part of their creation of selection tests used for medical training. Figure 2 illustrates the three phases and nine steps of the development process. In Phase 1, we developed the target attributes on which the content (scenarios and responses) of the SJT were based. We used an inductive approach with data gathered through observation of practising teachers, individual and focus group interviews with teachers and teacher educators, and questionnaires with teachers and teacher educators. An inductive approach to SJT development has been widely used in organizational psychology (Campion et al., 2014) and for developing selection tools for medical education (Patterson et al., 2016). In Phase 2, we created scenarios and responses for the SJT. In Phase 3, we carried out an initial validation of the SJT using concurrent data from current selection processes with participants from three ITE programs in the UK.

**Steps 1-3: Identifying target attributes.** Three steps were carried out to establish the target attributes for the SJT[1]. Defining the target attributes is an important step in developing SJTs, since creation of SJT content (scenarios and response options) is grounded in the target attributes. Step 1 consisted of full-day observations and in-depth interviews with two practising teachers in two schools. Step 1 was designed to provide an initial awareness of the activities and behaviors of the target teachers, inside and outside of the classroom. One teacher was a mid-career teacher and one was a newly-qualified teacher in her first year of practice after completing a teacher education program. A detailed summary report was produced describing the teachers' routines from the start of the day (e.g., 'up at 5 a.m., drive to gym') to the close of the day (e.g., 'as soon as child in bed, marking for 1 hour'). The purpose of Step 1 was not to provide an exhaustive or representative exploration of school life, but to

[1] Steps 1-3 were carried out for the development of an earlier version (for primary and secondary ITE applicants) of the SJT (see Klassen, Durksen, Rowett, & Patterson, 2014). In Step 4 we revised the target attributes created in Steps 1-3.

(re)familiarise the research team with the daily activities of teachers and the general functioning of schools.

In Steps 2 and 3, three focus group interviews were conducted in two schools ($n$ = 18) and one university teacher education program ($n$ = 10), and included practising teachers, school leaders, and teacher educators. Step 2 was designed to inductively identify the target attributes needed for successful novice teaching. The 28 expert participants were recommended by teacher education leaders and recruited from the pool of teachers and teacher educators who were involved in pre-service teacher supervision. We generated discussion using a critical incident approach where participants were encouraged to consider 'critical incidents' that led to positive or negative outcomes, e.g., *Think of a event where a newly-qualified teacher showed good (bad) judgment.* In addition, focus group participants were asked to generate and rate academic and non-academic attributes necessary for success for new teachers. Focus group data were collected and analysed using a content analysis approach. The focus group meetings resulted in the generation of 13 initial attributes (e.g., *caring, fairness, enthusiasm, reflection*) with behavioral descriptors.

Step 3 consisted of an iterative process of data reduction and integration led by three of the authors, and carried out through discussions with teachers and teacher educators about the importance of the 13 initial attributes (i.e., *How important are these attributes for new teachers?*). We used a multi-method consensus approach that integrated numerical ratings of the attributes with individual and group discussion of the relative importance of the attributes. In particular, we used a data reduction process that involved proposing clusters of domains to teacher and teacher educator focus groups and that asked *Which of these attributes are critical for the success in the teacher education program?* and *Which attributes are critical for the success of new teachers?* The 13 initial attributes were discussed individually and summarized into themes, or domains, with operational descriptors generated through discussion.

**Phase 1: Establishing Target Attributes**

After completion of the data reduction process, three composite domains-each consisting of two target attributes-emerged through further discussion and group consensus: *Empathy and Communication, Organisation and Planning, and Resilience and Adaptability*. The three composite domains were next evaluated for suitability to capture the key attributes specifically needed for novice teachers working in primary school contexts.

**Step 4: Reviewing target attributes.** Step 4 was conducted to evaluate and revise the target attributes specifically for the primary school environment. We posed three questions to seven experienced teacher educators from three UK university-based teacher education programs:

- *Do the three broad domains (and six target attributes) capture the non-academic attributes necessary for successful novice teaching at the primary school level?*
- *Are there any additional attributes that need considering?*
- *How do these attributes need adapting for a primary school teaching context?*

The review of target attributes resulted in retention of the three composite domains, but with a revision of the operational descriptors for a primary school environment. For example, the domain "Organisation and Planning" was broadened by consensus to include elements relating to managing competing priorities in order to capture the multiple demands primary school teachers face. Table 1 presents the three composite domains with the six target attributes and their descriptors. The domains generated in Steps 1-4 formed the foundation of the SJT content, and served as the basis for creating items (scenarios) and responses.

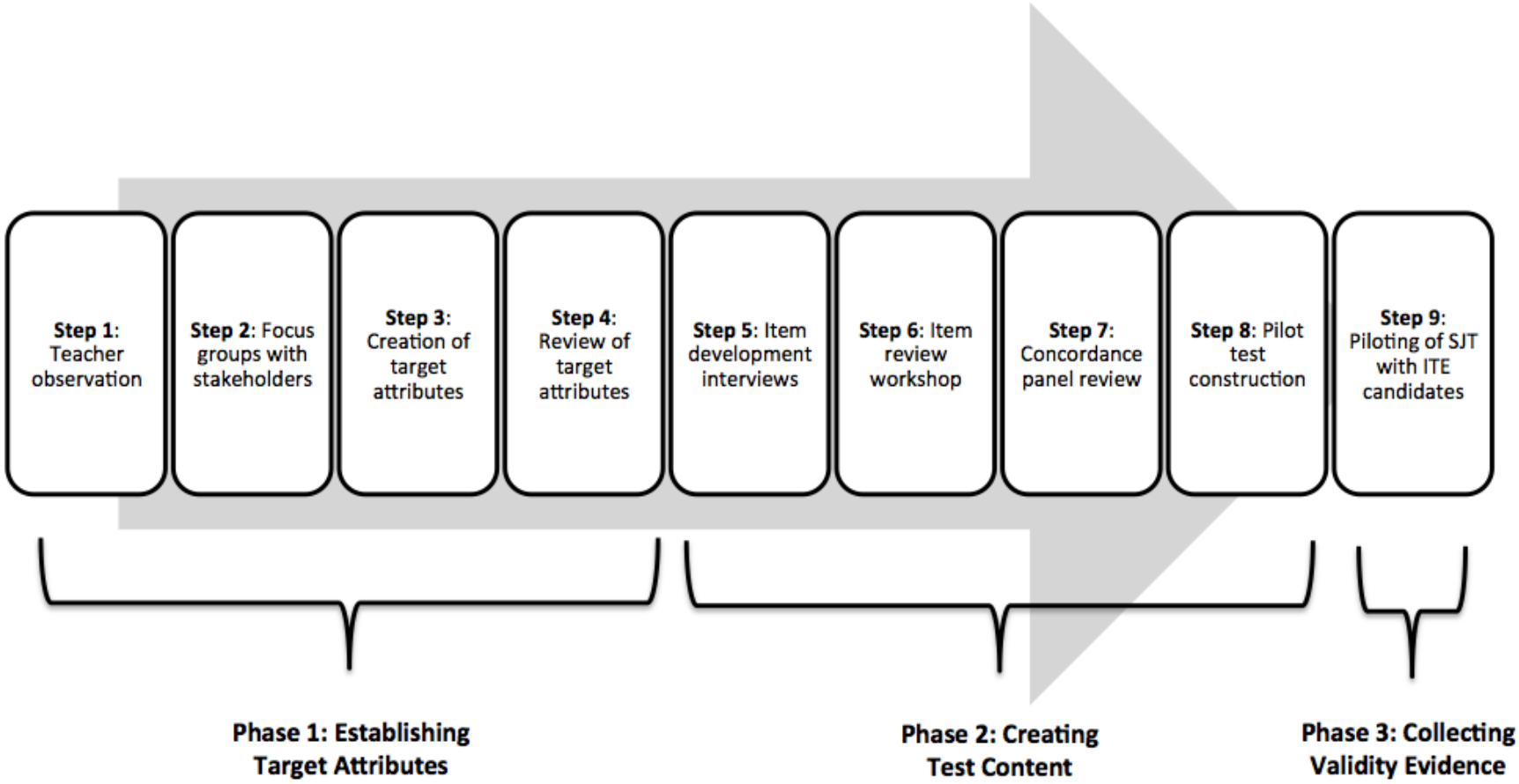



**Figure 2.** Nine steps of development of target attributes and pilot situational judgment test.

**Table 1.** Composite Domains and Target Attributes Identified for Teacher Selection SJT

| **Domain** | **Description** |
|---|---|
| *Empathy and Communication* | Candidate demonstrates active listening, and engages in an open dialogue with both pupils and colleagues. Candidate seeks advice pro-actively and is responsive to both professional feedback and pupils' needs. Candidate has the ability to adapt the style of communication and nature of dialogue appropriately. |
| *Organisation and Planning* | Candidate has the ability to manage competing priorities and display time management and personal organisation skills effectively, using these skills to enhance positive learning interactions with pupils. |
| *Resilience and Adaptability* | Candidate demonstrates the capability to remain resilient under pressure. Demonstrates adaptability, and an ability to change lessons (and the sequence of lessons) accordingly where required. Candidate has an awareness of their own level of competence and the confidence to either seek assistance, or make decisions independently, as appropriate. Is comfortable with challenges to own knowledge and is not disabled by constructive, critical feedback. Uses effective coping strategies. |

**Phase 2: Creating Test Content**

Phase 2 consisted of four steps (Steps 5 to 8) aimed at developing content for the SJT based on the target attributes.

**Step 5: Item development interviews.** Step 5 was conducted by trained interviewers (from an organizational behavior consulting firm) with practising teachers to develop scenarios and responses based on the identified target attributes. Eleven teachers who had experience working with novice teachers (i.e., as mentors of newly-qualified teachers) were individually interviewed in order to generate classroom scenarios and response options. A critical incidents method was used, whereby participants were asked to reflect on challenging situations that they had experienced as novice teachers or that they had observed when supervising novice teachers (Anderson & Wilson, 1997). Participants were guided to generate critical incidents related to the six target attributes. The resulting critical incidents were used as the basis for creating 54 SJT scenarios and responses. Table 2 presents an example SJT item that resulted from an item development interview.

**Step 6: Item review workshop.** A one-day workshop with eight experienced teachers from six UK primary schools (chosen for their involvement in supervising novice teachers), together with three teacher educators was held to review the 54 items (scenarios with associated response options) generated in Step 5. The workshop began with an introduction to item review principles and SJT attributes (e.g., *Is the item set in the correct context? Is the item set at an appropriate level for a novice teacher [not an experienced teacher]? Are the responses plausible? Does the content depend on specific knowledge* [which would unfairly discriminate against participants without a particular background]*?*). Participants were then arranged in pairs to review the 54 SJT items, followed by group work to revise problematic items. The workshop concluded with a calibration session where participants reviewed and discussed decisions made about content revision. The workshop resulted in an initial draft SJT consisting of all 54 items that were generated through item development interviews.

**Step 7: Concordance panel review.** In a concordance panel, test items are completed and evaluated by experts, and a scoring key is determined from a consensus of the experts (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). A concordance panel review session was conducted to identify a level of scoring consensus between expert reviewers in order to conclude which items had the highest degree of scoring agreement and to establish a scoring key. The 11 participants in the concordance panel were 9 experienced teachers and 2 teacher educators who worked closely with trainee teachers in schools and teacher education programs. Panel members completed the SJT in a 2-hour session, and provided additional feedback on the suitability and relevance of the scenarios and response options. Based on the scoring consensus and feedback on the 54 items, 35 items were selected for piloting with ITE candidates.

**Step 8: Pilot test construction.** The items were further revised based on feedback from the concordance panel (Step 7) and piloted with its scoring key. The pilot version of the SJT consisted of 35 scenarios designed for ITE candidates to complete in one hour. Five items represented the Organisation and Planning composite domain, 12 items represented Empathy and Communication, and 18 items represented Resilience and Adaptability. In order to reduce potential coaching effects (e.g., Whetzel & McDaniel, 2009), we used two response formats: 22 items used a ranking format (i.e., *Rank responses to this situation in order of appropriateness*) using a 5-point scale, and 13 items used a multiple response format (e.g., *Choose the three most appropriate actions to take in this situation*). Test scoring used a near

miss scoring approach: for ranking items, candidates received partial points for correct responses that were not in the optimal order. For example, four points were awarded to an item in correct position, three points for an item adjacent to correct position, two points for an item two positions away, and so on. For multiple response items, candidates received four points for each correct answer, giving a possible total of 12 points for each scenario.

**Table 2.** Example of SJT Scenario

| |
|---|
| You are teaching a lesson and have asked the students to individually complete an exercise that requires them to write down their responses. You have explained the exercise to the students and answered all of the questions that they have asked. As the students begin writing, one student, Ruby, starts to throw paper around and is clearly distracting the students sitting nearby. You know from previous incidents that Ruby often becomes frustrated when she does not understand how to complete activities, and that she often displays her frustration by being disruptive. |
| ***Choose the three most appropriate actions to take in this situation (*alternatively*, Rank the items in the most appropriate order)*** |
| • Send Ruby out the class if she continues to be disruptive |
| • Ask Ruby if she understands what the activity requires her to do |
| • Check in five minutes to see if Ruby has made progress with the exercise |
| • Tell Ruby that you are disappointed in her behavior |
| • Ask Ruby's classmate to discreetly provide help |
| • Stop the exercise and discuss the classroom behavior plan with the whole class |
| • *etc*. (eight total response options) |

*Note.* This is an example only, and is adapted from an item from the primary SJT.

## Phase 3: Collecting Reliability and Validity Evidence

**Step 9: Piloting of SJT with ITE candidates.** The final step in the last phase of development consisted of piloting the SJT with participants at two UK university ITE programs during their interview day. Participants were volunteers who were asked during the interview day if they would be willing to spend one hour completing the SJT. Interview day administrators estimated that 60% of candidates volunteered to complete the SJT during the course of the interview day, which consisted of procedures such as group activities, a written task, and individual interviews. A total of 124 candidates agreed to complete the SJT. Most of the candidates were female (81%) and white British (97.5 %), with a mean age of 22.3 years (range 20-34 years).

***Descriptive statistics.*** Analysis of the 35-item test scoring resulted in three items being dropped due to low item quality (low correlations with total test score), leaving 32 items for further analysis. The mean score of the test was 407.3 (*SD* = 33.19), with a range of 270 to 458. The difficulty level of the test was 76% (i.e., the mean score was 76% of the total possible score. As is conventional for SJTs, we did not calculate means, reliability coefficients, or validity coefficients for the individual domains (e.g., Lievens et al., 2008).

The reliability of the 32-item SJT ($\alpha$ = .79) compares favourably with other SJTs used in selection contexts (Whetzel & McDaniel, 2009). The mean test score was 407.3 (range 270 to 458) with a maximum possible score of 536. The distribution of the scores was near normal,

with a slight negative skew, meaning that most candidates scored in the higher range of the test rather than the lower range.

***Validity.*** We used interview scores for 108 participants provided by ITE program coordinators to test the SJT's concurrent validity. The seven scoring categories for the interview (scored on a 1-4 scale) were:

(1) ability to communicate in standard English
(2) pedagogical and subject knowledge
(3) reflections on experience
(4) understanding of education practice
(5) quality of thinking
(6) personal attributes and skills, and
(7) overall interview score.

Table 3 provides the means and standard deviations for the seven interview scores, and the correlations between the interview scores and total SJT score. The SJT showed significant positive correlations with each mean interview score ($.21 \leq r \leq .29$, $p < .01$), suggesting that the SJTs measured attributes that overlapped with the attributes measured by a wide range of interview indicators. The SJT showed a correlation of .29 with the overall interview score.

***Candidate reactions*.** We also collected data on candidates' perceptions of fairness, feasibility, and reasonableness of using SJT as part of the selection process because candidates' perceptions of the selection process influence their opinions of the organisation (Walker et al., 2013). From a recruitment perspective, a teacher training program's ability to successfully recruit applicants is influenced by the perceptions of current and past applicants, who may share word-of-mouth accounts about the fairness of the selection process, ultimately influencing the success of recruiting the best possible candidates.

Candidates reported a range of test completion times, with 56% of candidates reporting a completion time of 40–60 minutes and 42% of candidates reporting a completion time of less than 40 minutes. Most candidates (79%) *agreed/strongly agreed* that the test was "clearly relevant for those applying for ITE", and 74% *agreed/strongly agreed* that the level of difficulty was appropriate for ITE candidates. A majority of candidates (76%) *agreed/strongly agreed* that the content of the SJT appeared to be fair. Given an opportunity for open-ended responses, candidates commented that the test was useful to "place themselves in real life situations" and "far more applicable to the type of teaching experienced in the classroom" compared to other selection tests that they had taken for admission into other ITE programs. A minority of candidates commented that the test was too long and that, in some scenarios, it was difficult to judge the appropriate responses in the absence of additional information.

**Table 3.** Correlations Between Interview Scores and SJT Total Score

| | Interview domains | | | | | | |
|---|---|---|---|---|---|---|---|
| | Ability to communicate | Pedagogical & subject knowledge | Reflections on experience | Understanding of education | Quality of thinking | Personal attributes and skills | Mean interview score |
| Mean (SD) | 3.16 (.63) | 2.55 (.83) | 2.65 (.89) | 2.66 (.89) | 2.67 (.92) | 2.92 (.88) | 2.77 (.70) |
| Correlations with SJT score | .24* | .31** | .21* | .21* | .21* | .21* | .29** |

**Table 3.** Correlations Between Interview Scores and SJT Total Score

| | Interview domains | | | | | | |
|---|---|---|---|---|---|---|---|
| | Ability to communicate | Pedagogical & subject knowledge | Reflections on experience | Understanding of education | Quality of thinking | Personal attributes and skills | Mean interview score |
| Mean (SD) | 3.16 (.63) | 2.55 (.83) | 2.65 (.89) | 2.66 (.89) | 2.67 (.92) | 2.92 (.88) | 2.77 (.70) |
| Correlations with SJT score | .24* | .31** | .21* | .21* | .21* | .21* | .29** |

*Note.* $N = 108$. *$p < .05$. **$p < .01$.

## 3. DISCUSSION

Developing evidence-supported ITE selection practices is one approach to improving system-wide educational outcomes. In this proof-of-concept study, we presented the development and initial validation of a test for selection into primary ITE programs. The novel contribution of this article is that we show, as far as we know, the development of the first SJT-based selection test for primary teacher education programs, and although the results are encouraging, they represent the first step of many in a move to develop an operational selection tool. The results from the study suggest that the SJT methodology shows potential for selection purposes, with evidence of reliability, validity, and a positive response (e.g., perceived fairness) from ITE candidates.

We examined four research questions in this study. In response to the first research question (*Can a robust set of target attributes be established?*), three target attribute clusters were developed from a systematic inductive approach and endorsed by a diverse group of teachers and teacher educators. The three domains derived from the inductive development process used in our research have corollaries in other conceptual models of teacher effectiveness and teacher-student interactions. Pianta and Hamre's CLASS framework (2009) proposes three domains—emotional supports, classroom organization, and instructional supports—that can be mapped on to at least two of the inductively-derived domains in our model. Our domain of Empathy and Communication shares common ground with Pianta and Hamre's *emotional supports*, especially with the dimensions of *teacher sensitivity* and *regard for student perspectives*. Our domain of Organisation and Planning shares commonalities with *classroom organization,* with its dimensions of *behavior management* and *instructional learning formats*. Models of teacher effectiveness developed by other researchers, e.g., the *self-regulation skills* and *motivational characteristics* from the work of Kunter, Kleickmann, Klusmann, and Richter (2013) also share aspects of the domains developed in our model.

The inductive approach that we used, involving practicing teachers and teacher educators, was rigorous, and the target attributes were shown to be robust. However, further work is needed to expand the target attributes to include theory-derived (deductive) attributes that have been associated with teaching effectiveness, such as personality (Rockoff, Jacob, & Kane, 2011) and self-efficacy (Klassen & Durksen, 2014).

Our second and third research questions pertained to the psychometric properties of the proof-of-concept SJT. The psychometric results were acceptable, with a high level of reliability, a near-normal distribution, and significant empirical associations with interview criteria. Internal consistency reliability coefficients for SJTs are often low, partly because contextualised items (scenarios) tend to be complex and measure multiple constructs, even when they are designed to assess a particular attribute (Patterson et al., 2015).

The concurrent validity coefficient of $r = .29$ with overall interview score is encouraging for a proof-of-concept study and it is in line with fully developed SJTs (Christian et al., 2010). Further research will be needed to establish incremental validity of the SJT (i.e., what the SJT adds to selection decisions over-and-above other selection measures) and further work is needed to explore the predictive validity of the test using reliable and valid measures of teaching effectiveness (e.g., Pianta & Hamre, 2009).

Our fourth research question (*Do candidates view the SJT as fair and as a feasible selection method?*) was answered by candidates' generally positive responses to completing the SJT during selection. Candidates' perceptions of selection practices influence acceptance decisions, likelihood of litigation based on perceived unfairness of acceptance policies, and the academic reputation of the selecting institution. Previous research has shown that contextualised selection methods (e.g., SJTs) are perceived as being fairer than non-contextualised methods (e.g., personality tests; Bauer & Truxillo, 2006). Further steps to increase transparency might include providing candidates with information about how the test was developed and validated, and how SJT scores would be integrated into the selection process (e.g., the amount of weight an SJT score would carry in the overall selection process).

### 3.1. How an SJT might be used for selection into ITE programs

For live selection, admissions committees could use the SJT test in two ways. First, the test could be used for initial screening of non-academic attributes before candidates are invited to an expensive and time-consuming assessment centre or face-to-face interview day. The scoring of the SJT provides an overall score that can be weighted along with other assessment criteria, such as academic records, letters of reference, and interview scores, to produce a screening cut-off score. Most ITE programs already screen for academic attributes (e.g., review of academic transcripts) before inviting applicants to interviews; the SJT could be offered on site or at invigilated test centres for screening of non-academic attributes. SJTs could also be used in place of interviews, providing an efficient, economical, and arguably more valid assessment of non-academic attributes. Finally, SJTs could be used in addition to (or in combination with) currently used measures of non-academic attributes (e.g., letters of reference, interviews) as an additional source of data for decision making that might provide improvement in predicting who would most likely be most effective teachers.

***Next Steps.*** The results from the proof-of-concept SJT for selection are encouraging, but more psychometric and conceptual work is needed before such a test could be used for 'live' selection. Further work includes the generation of more SJT items to populate an item bank. Item development is an expensive and time-consuming process that requires item-writers to interview experienced teachers (who have worked with novice teachers) about critical incidents in a teaching context. Nevertheless, it is important to create a larger pool of validated items to populate alternate test forms in order to combat coaching effects (Whetzel & McDaniel, 2009).

The current study showed evidence of concurrent validity, but predictive validity evidence is needed to provide additional information about the usefulness of the SJT for ITE selection. While there is a lack of predictive validity research for any teacher selection process (Goldhaber, Grout, & Huntington-Klein, 2014), most SJT research explores concurrent, not predictive validity (Campion et al., 2014). A next step in developing a wider evidence base will be to study the relationships between pre-service teacher's SJT scores at entry and at the end of the ITE program. Further research will examine the longer-term predictive validity of SJTs using measures of teacher effectiveness in professional practice. Such tools may include the CLASS observation system (Pianta & Hamre, 2009), which involves observations of teachers' classroom behaviors, and the Tripod Survey, which involves anonymous student ratings of teacher-student interaction quality and classroom climate, which was used in the *Measures of Effective Teaching* project (Kane & Staiger, 2012). CLASS and Tripod measures are well-researched teacher effectiveness tools that have been rigorously validated over the last decade.

A further step will be to examine the relative effectiveness of competing constructs and selection measures. Lievens & Patterson (2011) used structural equation modelling to estimate the relative influence of SJTs alongside two other variables in predicting supervisor ratings of medical trainees' performance. Results showed that all three variables were valid predictors of job performance, with SJTs showing incremental validity over the academic measures. Final validation of an SJT designed for ITE selection would test incremental validity over the academic and non-academic measures currently used for selection.

We used a bottom-up inductive approach by way of a critical incidents technique to develop the target attributes to base our test content on. Another approach used in SJT research is a theory-based or deductive approach (Campion et al., 2014), in which target attributes are based on existing theoretical models such as personality and motivation. Our research team is currently developing theory-based SJTs to assess motivation (e.g., self-efficacy) and personality as target attributes.

### 3.2. International research

Interest in developing evidence-led ITE selection methods is not unique to the UK, and research on identifying key factors related to success in ITE programs is being carried out in a range of international settings. One key question in our international projects on ITE selection is the extent to which teaching attributes identified in one context are endorsed in another national context. A key principle in developing selection methods internationally is to recognize that although some attributes of effective teachers may be universal, other attributes measured need to reflect local contexts (Lievens et al., 2015).

### 3.3. Limitations

The sample of participants in Step 9 (pilot study) was smaller than anticipated and less ethnically diverse than the overall population of teachers in the UK (97.5% White British in our sample versus 93% nationally). However, the gender balance of participants in our study was the same (80%) as the gender balance reported for teachers nationally (Department for Education, 2016). One stated advantage of using SJTs for selection—that they are less prone to inter-group differences than other selection methods such as cognitive tests (Whetzel & McDaniel, 2009)—was not tested in this study, and more diverse samples will be needed to establish inter-group profiles to further investigate the fairness of SJTs.

## 4. CONCLUSIONS

This study is the first to report the development of a proof-of-concept SJT to select candidates into ITE programs. The results should be interpreted cautiously, with a restricted sample involving concurrent validity data. A selection system needs to be robust, transparent, and perceived as fair by applicants, and built on evidence collected from multiple methods. In many contexts, cost-effectiveness is also an important factor in choosing selection tools: an SJT can be used as a screening tool to evaluate non-academic attributes alongside evaluation of academic attributes, thus reducing the time and cost involved in the selection process. In settings where large numbers of candidates apply for limited spaces, SJTs could be used in conjunction with other data (such as academic records) to select a reduced number of candidates for more intensive selection procedures such as face-to-face interviews. The intention of this proof-of-concept study was to show the feasibility of developing an SJT for selection into teacher education programs, but exactly how, when, and the extent to which this method might be used would be determined by local contexts and needs.

**Funding**

This work was supported by research funding from the European Research Council.

## 5. REFERENCES

Anderson, L. & Wilson, S. (1997) Critical incident technique. In D.L. Whetzel and G.R. Wheaton (eds.), *Applied Measurement Methods in Industrial Psychology*. Palo Alto, CA: Davies-Black.

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Improvement in early career teacher effectiveness. *AERA Open*, 1(4), 1-23.

Barber, M. & Mourshed, M. (2007). *How the world's best performing school systems come out on top*. London: McKinsey & Company.

Bauer, T. N., & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J. A. Weekley & R. E. Ployhart (Eds.) *Situational judgment tests: Theory, measurement, and application* (pp. 233-249). Mahwah, NJ: Lawrence Erlbaum Associates.

Bergman, M. E., Drasgow, F., Donovan, M. a., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14, 223-235.

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, 27, 283–310.

Casey, C. E., & Childs, R. A. (2007). Teacher education program admission criteria and what beginning teachers need to know to be successful teachers. *Canadian Journal of Educational Administration and Policy*, 67, 1-24.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63, 83-117.

Department for Education (2016). Statistics at DfE. Retrieved from: https://www.gov.uk/government/organisations/department-for-education/about/statistics

Elliott, J. G., Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., & Hoffman, N. (2011). The socially skilled teacher and the development of tacit knowledge. *British Educational Research Journal*, 37, 88-103.

Goldhaber, D., Grout, C., & Huntington-Klein, N. (2014). *Screen twice, cut once: Assessing the predictive validity of teacher selection tools*. CEDR Working Paper No. 2014-9: University of Washington, Seattle, WA.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4-27.

Hanushek, E. A., & Rivkin, S. G. (2011). The Distribution of Teacher Quality and Implications for Policy. *Annual Review of Economics*, 4, 131-157.

Hobson, A. J., Ashby, P., McIntyre, J., & Malderez, A. (2010). *International approaches to teacher selection and recruitment*. OECD Education Working Paper No.47: Organisation for Economic Co-operation and Development.

Hooper, A. C., Jackson, H. L., & Motowidlo, S. J. (2004). *Situational judgment measures of*

*personality and work-relevant performance*. Paper presented at the 112th annual meeting of the American Psychological Association, Honolulu, HI.

Johnson, R. E., & Saboe, K. N. (2011). Measuring implicit traits in organizational research: Development of an indirect measure of employee implicit self-concept. *Organizational Research Methods*, 14, 530-547.

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Research Paper. MET Project. Bill & Melinda Gates Foundation.

Klassen, R. M., & Dolan, R. (2015, September). *Selection for teacher education in the UK and the Republic of Ireland: A proposal for innovation*. Presented at the meeting of the European Conference on Educational Research, Budapest, Hungary.

Klassen, R. M., & Durksen, T. L. (2014). Weekly self-efficacy and work stress during the final teaching practicum: A mixed methods study. *Learning and Instruction*, 33, 158-169.

Klassen, R.M., Durksen, T.L., Rowett, E., & Patterson, F. (2014). Applicant reactions to a situational judgment test used for selection into initial teacher training. *International Journal of Educational Psychology*, 3, 104-125.

Knight, F. B. (1922). Qualities related to success in elementary school teaching. *The Journal of Educational Research*, 5, 207-216.

Kunter, M., Kleickmann, T., Klusmann, U., & Richter, D. (2013). The development of teachers' professional competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 63-77). New York, NY: Springer.

Lievens, F., Corstjens, J., Sorrel, M. A., Abad, F. J., Olea, J., & Ponsoda, V. (2015). The cross-cultural transportability of situational judgment tests: How does a US-based integrity situational judgment test fare in Spain. *International Journal of Selection and Assessment*, 23, 361-372.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology*, 96, 927-940.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, 37, 426-441.

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic succss and job performance. *Journal of Applied Psychology*, 97, 460-468.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology*, 60, 63-91.

McDaniel, M. A., Whetzel, D., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79, 599-616.

Mikitovics, A., & Crehan, K. D. (2002). Pre-professional skills test scores as college of education admissions criteria. *The Journal of Educational Research*, 95, 215-223.

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit

trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology*, 95, 321-333.

Patterson, F., Ashworth, V., Mehra, S., & Falcon, H. (2012). Could situational judgement tests be used for selection into dental foundation training? *British Dental Journal*, 213, 23–26.

Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice. AMEE Guide No. 100. *Medical Teacher*, 38, 3-17.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education and training? Evidence from a systematic review. Medical Education, 50, 36–60.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38, 109-119.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education*, 6, 43-74.

Sahlberg, P. (2014). *Finnish lessons 2.0: What can the world learn from educational change in Finland?* New York, NY: Teachers College Press.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135-155). Mahwah, NJ: Erlbaum.

Sclafani, S. K. (2015). Singapore chooses teachers carefully. Phi Delta Kappan, 97(3), 8-13.

Shultz, M. M., & Zedeck, S. (2012). Admission to Law School: New Measures. *Educational Psychologist*, 47, 51-65.

Staiger, D. O., & Kane, T. J. (2015). Making Decisions with Imprecise Performance Measures. In T. Kane, K. Kerr, R. Pianta, & Bill and Melinda Gates Foundation (Eds.), *Designing Teacher Evaluation Systems* (pp. 144-169). San Francisco, CA: Jossey-Bass.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, 24(3), 97-118.

Thomson, D., Cummings, E., Ferguson, A. K., Moizumi, E. M., Sher, Y., Wang, X., Broad, K., & Childs, R. A. (2011). A role for research in initial teacher education admissions: A case study from one Canadian university. *Canadian Journal of Educational Administration and Policy*, 67, 1–24

Walker, H. J., Bauer, T. N., Cole, M. S., Bernerth, J. B., Feild, H. S., & Short, J. C. (2013). Is this how I will be treated? Reducing uncertainty through recruitment interactions. *Academy of Management Journal*, 56, 1325-1347.

Watt, H. M. G., Richardson, P. W., & Wilkins, K. (2014). Profiles of professional engagement and career development aspirations among USA preservice teachers. *International Journal of Educational Research*, 65, 23-40.

Whetzel, D., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, 19, 188-202.

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 115-121*

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

*"Research Article"*

# Pre-service Elementary Teachers' Attitudes Towards Chemistry Course

**Seçil ERÖKTEN***

Pamukkale University, Faculty of Education, Science Education Dept., Denizli, Turkey

***Abstract***

The current research aims to investigate the attitudes of pre-service elementary teachers towards chemistry course according to their gender and the type of high school that they graduated from. As a scanning model, the current study was conducted in spring semester of 2015-2016 academic years with the students of Elementary Education Department at Faculty of Education in Pamukkale University. 99 students who were attending the chemistry course in their first year curriculum participated in the study. The data collection tool was "Attitudes Towards Chemistry Course Scale" which was developed by Hançer, Uludağ and Yılmaz (2007). Whether the attitude scores of participants changed according to gender and the high school that they graduated from were tested with independent samples t-test and one-way ANOVA. According to the analyses, the students were observed to have negative attitudes towards chemistry class and these negative attitudes did vary according to neither gender nor the high school that graduated from.

***Article Info***

***Received***
*21 March 2016*

***Revised***
*24 October 2016*

***Accepted***
*11 December 2016*

***Key words***
*Attitude, Chemistry Class, Classroom Teaching, School Type, Gender*

# Sınıf Öğretmenliği Öğrencilerinin Kimya Dersine Yönelik Tutumları

***Özet***

Bu araştırma, Sınıf Öğretmenliği öğrencilerinin kimya dersine yönelik tutumlarının cinsiyet ve mezun oldukları okul türleriyle ilişkisini incelemeyi amaçlamaktadır. Tarama modelinde bir araştırmadır. Araştırma; 2015-2016 eğitim-öğretim yılı bahar döneminde, Pamukkale Üniversitesi, Eğitim Fakültesi, Sınıf Öğretmenliği Anabilim Dalı öğrencileriyle gerçekleştirilmiştir. Araştırmaya 1. sınıf ders programında yer alan kimya dersine devam eden 99 öğrenci katılmıştır. Veri toplama aracı olarak Hançer, Uludağ ve Yılmaz (2007) tarafından geliştirilen "Kimya Dersine Yönelik Tutum Ölçeği" kullanılmıştır. Sınıf öğretmenliği öğrencilerinin tutum puanlarının cinsiyet ve mezun oldukları okul türlerine göre farklılık gösterip göstermediği bağımsız örneklem t-testi ve tek yönlü varyans analizi ile incelenmiştir. Verilerin analizi sonucunda öğrencilerin kimya dersine yönelik olumsuz tutuma sahip oldukları, bu olumsuz tutumlarının cinsiyet ve mezun oldukları okul türüne göre de değişmediği gözlenmiştir.

***Makale Bilgisi***

***Gönderildi***
*21 Mart 2016*

***Düzeltildi***
*24 Ekim 2016*

***Kabul Edildi***
*11 Aralık 2016*

***Anahtar Kelimeler***
*Tutum, Kimya Dersi, Sınıf Öğretmenliği, Okul Türü, Cinsiyet*

*Author Phone: +90 258 296 1176 E-mail: erokten@pau.edu.tr

2148-7456 / © 2016 DOI: 10.21449/ijate.275803

## 1. GİRİŞ

Uluslararası alanda bir ülkenin yerini, o ülkedeki bilginin kalitesi ve iyi yetişmiş insan gücü belirler (Hançer, 2005). Teknolojik değişme ve gelişme fen bilimleri sayesinde gerçekleşir. Fen bilimleri, bilimin ve teknolojinin gelişmesinde çok önemli bir yere sahiptir. Bu nedenle günümüzde fen bilimleri eğitiminin önemi artmaktadır (Demirci, 1993). Fen bilimlerinin temelini fizik, kimya ve biyoloji oluşturmaktadır. Kimya, maddenin yapısını, maddenin özelliklerini ve maddelerin birbirleriyle ilişkilerini araştırmaktadır. Kimya biliminin incelediği konu alanı çok geniştir. Canlı ve cansız varlıkların yapısından çevre sorunlarına kadar günümüzde yaşanan birçok olayı incelemektedir. Dünyayı tanımlamada, doğada gerçekleşen olayları açıklamada, doğa olaylarında neden sonuç ilişkisi kurmada kimyadan yararlanılır. Bu nedenle yaşamımızın devamı için kimya bilimine, dolayısıyla da kimya eğitimine önem vermemiz gerekmektedir. Kimya eğitiminde, bireylerin keşfederek bilgiye kendilerinin ulaşması, yeni bilgilere ulaştıkça dünyaya bakışını revize etmesi ve öğrenme hevesinin gelişmesi çok önemlidir (Sezgin Saf, 2011). Kimya eğitiminde öğrencilerin; dünyayı anlamaları, anlamlı sorular sorup, gözlem ve deney yapıp, analiz etmeleri, sorumluluklarının bilincinde ve bilgisinde olabilmeleri için kimya dersine karşı olumlu tutum geliştirmeleri gerekmektedir (İnce Aka ve Sert Çıbık, 2015). Öğrencilerin bir derse yönelik tutumları; dersle ilgili olumlu düşüncelere sahip olmaları, dersi sevmeleri ya da derse karşı olumsuz düşüncelere sahip olmaları, dersi sevmemeleri şeklinde ifade edilebilir (İnce Aka ve Sarıkaya, 2014). Anderson (1988), özel bir durumla karşılaşıldığında, uygun olan ve olmayan tarzda tepki vermek için bireyin eğilimli olmasını ya da hazırlanmasını sağlayan, orta düzeyde yoğunluğu olan bir heyecan olarak tutumu tanımlamaktadır. Bir derse karşı olumlu tutum; derse katılma, derste soru sorma, sorulara cevap verme ve bundan zevk alma gibi davranışlar şeklinde gözlenebilir (Özçelik, 1998). Öğrencilerdeki mevcut tutumun belirlenmesi; gelecekteki davranışları hakkında fikir sahibi olmayı sağlayacak ve ulaşılması istenilen değişikliklerin gerçekleştirilmesine yardımcı olacaktır (Nuhoğlu, 2008). Eğitimde istenilen başarının elde edilebilmesi için öğrencilerin tutumlarının bilinmesi gerekmektedir. Bu nedenle öğrenci tutumlarının belirlenmesi önemli hale gelmiştir (Meyveci, 1997). Tutum ile akademik başarı arasında pozitif yönde bir ilişki olduğu söylenebilir. Öğrencilerin derse karşı tutumları olumlu ise o derse ilişkin akademik başarıları da yüksek olacaktır (Sezgin Saf, 2011; Karasakaloğlu ve Saracaloğlu, 2009). Cheung (2009), öğrencilerin kimya dersine karşı tutumlarının akademik başarıları etkileyen bir değişken olduğunu belirtmiştir. Bennet, Rollnick, Green ve White (2001) yaptıkları çalışmada, kimyaya karşı tutumun akademik başarıya etkisini incelenmişler ve olumsuz tutuma sahip öğrencilerin akademik başarılarının da düşük olduğunu belirlemişlerdir. Salta ve Tzougraki (2004) 11. Sınıf öğrencilerinin kimya dersine karşı tutumlarının cinsiyete göre değişip değişmediğini incelemiş, erkek öğrenciler lehine anlamlı bir fark bulmuştur. Yılmaz (2007) yabancı dil öğreniminde motivasyonun önemini araştırmış, okul türü ve cinsiyet açısından farklılık gösterip göstermediğini incelemiştir. Motivasyon düzeylerinin okul tipine göre değiştiğini ve cinsiyetler açısından yaptığı karşılaştırmada da kız öğrenciler lehine anlamlı bir fark tespit etmiştir. Kıngır ve Yazıcı (2007) lise öğrencilerinin kimya dersine karşı tutumlarını sosyoekonomik durumları, cinsiyetleri ve okul türleri açısından incelemiş, cinsiyet ve sosyoekonomik durum açısından bir farklılık olmadığını, okul türüne göre anlamlı bir farklılık olduğunu tespit etmiştir.

Kimya eğitiminde istenilen sonuca ulaşabilmek için öncelikle öğrencilerin tutumlarının ölçülmesi gerekmektedir. Gelecek yeni nesle Sınıf Öğretmenleri şekil vereceği için Sınıf Öğretmenliği öğretmen adaylarının kimyaya yönelik tutumlarının ne olduğunu tespit etmek gerekmektedir. Bu araştırmada, Sınıf Öğretmenliği öğrencilerinin kimyaya yönelik tutumlarını tespit etmek, cinsiyet ve mezun oldukları okul türleri arasında farklılık olup olmadığını belirlemek amaçlanmıştır. Bu amaç doğrultusunda aşağıdaki sorulara yanıt aranmıştır.

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 115-121*

1. Sınıf Öğretmenliği öğrencilerin kimyaya yönelik tutumları ile cinsiyetleri arasında ilişki var mıdır?
2. Sınıf Öğretmenliği öğrencilerin kimyaya yönelik tutumları ile mezun oldukları okul türleri arasında ilişki var mıdır?

## 2. YÖNTEM

Tarama modelinde bir araştırmadır. Araştırma; 2015-2016 eğitim-öğretim yılı bahar döneminde gerçekleştirilmiştir.

### 2.1. Evren ve örneklem

Sınıf Öğretmenliği 1. sınıf öğrencileri çalışmanın evrenini oluşturmaktadır. Çalışma grubu ise Pamukkale Üniversitesi, Eğitim Fakültesi, Sınıf Öğretmenliği Anabilim Dalı öğrencileridir. Araştırmaya 1. sınıf ders programında yer alan kimya dersine devam eden 99 öğrenci katılmıştır.

### 2.2. Veri Toplama Aracı

Veri toplama aracı olarak Hançer, Uludağ ve Yılmaz (2007) tarafından geliştirilen "Kimya Dersine Yönelik Tutum Ölçeği" kullanılmıştır. Ölçek 16 tane olumlu, 16 tane olumsuz 32 maddeden oluşmaktadır. Araştırmacılar testin güvenlik katsayısını Cronbach Alpha = 0,87 olarak belirlemişlerdir. Ölçekten elde edilebilecek en yüksek puan olan 160 olumlu tutumları, en düşük puan olan 32 olumsuz tutumları, 96 puan ise nötr tutumları göstermektedir. Bu durumda; 96'ın üzerindeki puanlar olumlu tutumu, altında kalan puanlar olumsuz tutumu göstermektedir.

### 2.3. Verilerin Analizi

Araştırmaya katılan öğrencilerin 80'i kız, 19'u erkek öğrencidir. Öğrencilerin 47'si Anadolu Lisesi, 21'i Düz Lise, 25'i Öğretmen Lisesi, 6'sı Meslek Lisesi mezunu olduğu tespit edilmiştir. Ayrıca 9 öğrenci daha önce kimya dersi gördüğünü, 90 öğrenci kimya dersi görmediğini belirtmiştir. Olumlu maddelerin seçenekleri 5'den 1'e kadar, olumsuz maddelerin seçenekleri 1'den 5'e kadar değerler verilerek ölçeğin kodlaması yapılmış ve araştırmanın güvenlik katsayısı Cronbach Alpha = 0,939 olarak hesaplanmıştır. Normal dağılım gösterip göstermediğini kontrol etmek amacıyla Kolmogorov-Smirnov analizi yapılmış ve normal dağılım gösterdiği belirlenmiştir ($Z=0,676$; $p>0,05$).

## 3. BULGULAR

Sınıf Öğretmenliği öğrencilerinin kimyaya yönelik tutumları tespit etmek için, cinsiyet ve mezun oldukları okul türleri arasında farklılık olup olmadığını belirlemek amacıyla istatistiksel analizler yapılmıştır.

### 3.1. Birinci Alt Probleme İlişkin Bulgular

1. alt problem cümlesi olan "Sınıf Öğretmenliği öğrencilerin kimyaya yönelik tutumları ile cinsiyetleri arasında ilişki var mıdır?" sorusuna cevap bulabilmek için bağımsız örneklem t-testi analizi yapılmıştır. Elde edilen sonuçlar Tablo 1'de verilmektedir.

Tablo 1'de de görüldüğü gibi kız ve erkek öğrencilerin kimyaya yönelik tutumlarında anlamlı bir farklılık tespit edilememiştir ($t_{97}=0,964$, $p>0,05$). Kız öğrencilerinin ortalamaları 90,91, erkek öğrencilerin ortalamaları 85,58 olarak bulunmuştur.

**Tablo 1.** Bağımsız örneklem *t*-testi Sonuçları

| | N | X | s | sd | *t* | p |
|---|---|---|---|---|---|---|
| Kız | 80 | 90,91 | 22,133 | 97 | 0,964 | 0,338 |
| Erkek | 19 | 85,58 | 19,599 | | | |

### 3.2. İkinci Alt Probleme İlişkin Bulgular

2. alt problem cümlesi olan "Sınıf Öğretmenliği öğrencilerin kimyaya yönelik tutumları ile mezun oldukları okul türleri arasında ilişki var mıdır?" sorusuna cevap bulabilmek için tek yönlü varyans analizi yapılmıştır. Elde edilen sonuçlar Tablo 2'de verilmektedir. Öğrencilerin mezun oldukları okul türleri arasında anlamlı bir farklılık tespit edilememiştir ($F(3,95)=2,482$, $p>0,05$).

**Tablo 2.** Tek yönlü varyans analizi sonuçları

| | Kareler top. | sd | Kareler ort. | *F* | p |
|---|---|---|---|---|---|
| Gruplar arası | 3347,024 | 3 | 1115,675 | 2,482 | 0,066 |
| Grup içi | 42704,754 | 95 | 449,524 | | |
| Toplam | 46051,778 | 98 | | | |

## 4. SONUÇ VE TARTIŞMA

Sınıf Öğretmenliği öğrencilerinin kimyaya yönelik tutumları tespit etmek, cinsiyet ve mezun oldukları okul türleri arasında farklılık olup olmadığını belirlemek amacıyla yapılan bu çalışmada elde edilen bulgulardan şu sonuçlara varılmıştır:

1. alt problem cümlesi olan "Sınıf Öğretmenliği öğrencilerin kimyaya yönelik tutumları ile cinsiyetleri arasında ilişki var mıdır?" sorusunun cevabını bulmak için yapılan bağımsız örneklem t-testi sonucunda anlamlı bir fark olmadığı tespit edilmiştir ($t97=0,964$, $p>0,05$). Kız öğrencilerinin ortalamaları 90,91, erkek öğrencilerin ortalamaları 85,58 olarak bulunmuştur. Kız ve erkek öğrencilerin ortalama puanlarının nötr puan olan 96 puanın altında olduğu görülmektedir. Elde edilen bu ortalamalardan da anlaşılacağı gibi ne kız öğrenciler ne de erkek öğrencilerin kimya dersine yönelik olumlu bir tutuma sahip olmadığı görülmektedir.

2. alt problem cümlesi olan "Sınıf Öğretmenliği öğrencilerin kimyaya yönelik tutumları ile mezun oldukları okul türleri arasında ilişki var mıdır?" sorusunun cevabını bulmak için yapılan tek yönlü varyans analizi sonucunda öğrencilerin mezun oldukları okul türleri arasında da bir fark bulunmadığı belirlenmiştir ($F(3,95)=2,482$, $p>0,05$). Bu durumda öğrencilerin kimyaya yönelik tutumlarında mezun oldukları okul türünün de etkisi olmadığı görülmektedir.

Öğrencilerin kimyaya yönelik tutumları 48 puanla 136 puan arasında değişmektedir. Öğrencilerin kimyaya yönelik tutumlarının genel olarak ortalaması 89,89 olarak bulunmuştur. Bulunan bu genel ortalama da nötr puan olan 96 puanın altında kalmaktadır. Bu puanda öğrencilerin genel olarak kimyaya yönelik olumlu tutuma sahip olmadığını göstermektedir. Ayrıca 90 öğrencinin daha önce kimya dersi görmediği ancak 9 öğrencinin daha önce kimya dersi gördüğü tespit edilmiştir. Bu 90 öğrenci eşdeğer ağırlıklı şubeden mezun olduklarını en son 9. sınıfta kimya dersi gördüğünü belirtmişlerdir. Sınıf Öğretmenliği öğrencilerinin kimyaya karşı olumsuz tutum sergilemelerinin nedeninin eşdeğer ağırlıklı şubeden mezun olmalarından kaynaklandığını söyleyebiliriz.

İnce Aka (2012), öğrencilerin kimya dersine yönelik ilgi ve tutumlarını incelediği çalışmada kız ve erkek öğrencilerin tutumları arasında bir fark tespit edememiştir. Araştırmamız İnce Aka'nın yaptığı çalışma ile benzerlik göstermektedir. Yapılan çalışmalar incelendiğinde; araştırmacılar öğrencilerin dersi sevmeleri ile başarı arasında doğru bir

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 115-121*

orantının olduğunu belirtmektedirler (Oral ve McGivney, 2011; Altınok, 2005; Şişman, Acat, Aypay ve Karadağ, 2011). Bu nedenle öğrencilerin kimya dersine başarılı olabilmeleri için kimyaya yönelik olumlu tutum içinde olmaları gerektiği düşünülmektedir. Daha sonraki çalışmalarda Sınıf Öğretmenliği öğrencilerin kimyaya yönelik tutumlarıyla kimya başarıları arasında bir ilişki olup olmadığı incelenebilir.

## 5. KAYNAKLAR


Altınok. H. (2005). Cinsiyet ve Başarı Durumlarına Göre İlköğretim 5. Sınıf Öğrencilerinin Fen Bilgisi Dersine Yönelik Tutumları, *Eurasian Journal of Educational Research*, 17, 81-91.

Anderson, L.W. (1988). *Attitudes and their Measurement*. In Keeves, J.P. (ed). *Educaitonal Reseach, Methodology and Measurement: an International Handbook*. New York: Pergamon Press.

Bennet, J., Rollnick, M., Green, G. and White, M. (2001). The Development and Use of an Instrument to Assess Students' Attitude to the Study of Chemistry, *International Journal of Scinece Education*, 23(8), 833-845.

Cheung, D. (2009). Students' Attitudes Twoard Chemistry Lessons: The Interaction Effect Between Grade Level and Gender, *Reseacrh in Science Educaiton*, 39, 75-91.

Demirci. B. (1993). Çağdaş Fen Bilimleri Eğitimi ve Eğitimcileri, H.Ü. Eğitim Fakültesi Dergisi, 9, 115-124.

Hançer, A.H, Uludağ, N. ve Yılmaz, A. (2007). Fen Bilgisi Öğretmen Adaylarının Kimya Dersine Yönelik Tutumlarının Çeşitli Değişkenlere Göre Değerlendirilmesi, *H.Ü. Eğitim Fakültesi Dergisi*, 32, 100-109.

Hançer, A.H. (2005). *Fen Eğitiminde Yapılandırmacı Yaklaşıma Dayalı Bilgisayar Destekli Öğrenmenin Öğrenme Ürünlerine Etkisi*, yüksek lisans tezi, G.Ü. Eğitim Bilimleri Enstitüsü, Ankara.

İnce Aka, E. (2012). *Asitler ve Bazlar Konusunun Öğretimde Kullanılan Probleme Dayalı Öğrenme Yönteminin Farklı Değişkenler Üzerine Etkisi ve Yönteme İlişkin Öğrenci Görüşleri*, doktora tezi, G.Ü. Eğitim Bilimleri Enstitüsü, Ankara

İnce Aka, E. ve Sarıkaya, M. (2014). Probleme Dayalı Öğrenme Yönteminin Fen Bilgisi Öğretmen Adaylarının Kimya Dersine Yönelik Tutumlarına Etkisi, *GEFAD/GUJGEF*, 34(3), 455-467.

İnce Aka, E. ve Sert Çıbık, A. (2015). Fen Bilgisi Öğretmen Adayların Kimya Dersine Yönelik Tutumlarının Farklı Değişkenlere Göre İncelenmesi, *GEFAD/GUJGEF*, 35(3), 557-573.

Karasakaloğlu, N. ve Saracaloğlu, A.S. (2009). Sınıf Öğretmeni Adaylarının Türkçe Dersine Yönelik Tutumları, Akademik Benlik Tasarımları ile Başarıları Arasındaki İlişki, *Yüzüncü Yıl Üniversitesi, Eğitim Fakültesi Dergisi*, 6(1), 343-363.

Kıngır, S. ve Yazıcı, N. (2007). *Lise 1 Öğrencilerinin Kimya Dersine İlişkin Tutumları ve Motivasyon Üzerine Araştırma*, 1. Ulusal Kimya Eğitimi Kongresi, İstanbul.

Meyveci, N. (1997). *Bilgisayar Destekli Fizik Öğretiminin Öğrenci Başarısına ve Öğrencinin Bilgisayara Yönelik Tutumuna Etkisi*, yüksek lisans tezi, A.Ü. Sosyal Bilimler Enstitüsü, Ankara.

Nuhoğlu, H. (2008). İlköğretim Fen ve Teknoloji Dersine Yönelik Bir Tutum Ölçeğinin Geliştirilmesi, *İlköğretim Online*, 7(3), 627-638.

Oral, I. ve McQivney, E. (2011). *Türkiye'de Matematik ve Fen Bilimleri Alanlarında Öğrenci Performansı ve Başarısının Belirleyicileri*: TIMSS 2011 Analizi, Eğitim Reformu Girişimi, 1-31 (13.03.2016 tarihinde erişimiştir).

Özçelik, D.A. (1998). Ölçme ve Değerlendirme, ÖSYM Yayınları, Ankara

Salta, K. and Tzougraki, C. (2004). Attitudes Toward Chemistry Among 11th Grade Students İn High Schools in Greece, *Science Education*, (88), 535547.

Sezgin Saf, A. (2011). *Ortaöğretim 9. Sınıf Öğrencilerinin Kimya Dersine İlişkin Tutum, Motivasyon ve Öz Yeterlik Algılarının Çeşitli Değişkenler ile İncelenmesi*, Yüksek lisans tezi, Selçuk Üniversitesi, Eğitim Bilimleri Enstitüsü, Konya.

Şişman, M., Acat, M.B., Aypay, A. ve Karadağ, E. (2011). *TIMSS 2007 Ulusal Matematik ve Fen Raporu 8. Sınıflar*, MEB, Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı, Hermes Ofset, Ankara.

Yılmaz, E. (2007). *Ortaöğretimde İngilizce Derslerinde Öğrenci Başarısına Motivasyonun Rolü (Bartın Örneği)*, Yüksek Lisans Tezi, Karaelmas Üniversitesi, Sosyal Bilimler Enstitüsü, Zonguldak

## Summary

### Introduction

Technology changes and develops by means of physical sciences. Physical sciences have a very important role in the development of science and technology. Thereby, the education of physical sciences is increasingly popular in our day and age (Demirci, 1993). Physics, chemistry and biology constitute the core of physical sciences. Among these, the field of chemistry investigates the structure of matter, the characteristics of matter and the interaction of matters with each other. In order to understand and continue our lives, we have to give importance to the science of chemistry, hence to the chemistry education. In chemistry education, it is very important for individuals to reach the knowledge on their own via exploring, to revise their point of views for the world as they reach the new knowledge, and to develop an eagerness to learn (Sezgin Saf, 2011). The attitudes of the students towards a course can be defined as their having positive or negative feelings about the course, and their liking and dislike for the course (İnce Aka and Sarıkaya, 2014). Determining the current attitudes of the students will lead to get an idea about their future behaviour and help to carry out the targeted changes (Nuhoğlu, 2008). In order to achieve the desirable educational outcomes, it is necessary to know the attitudes of students. It can also be said that there is a positive relationship between attitudes and academic achievement.

In order to reach the intended outcomes for the chemistry education, students' attitudes should be firstly measured. It is primarily necessary to determine the attitudes of pre-service elementary teachers because they are responsible to shape the framework of new generations. In this sense, the aim of the current study was to determine the pre-service elementary teachers' attitudes towards chemistry and to determine whether gender and the type of high school that they graduated from affected the outcome. In regard of these aims, the current study tried to answer the following questions:

1. Is there a relationship between the attitudes of pre-service elementary teachers towards the chemistry course and their gender?

2. Is there a relationship between the attitudes of pre-service elementary teachers towards the chemistry course and the high school type that they graduated from?

### Methodology

Survey method was adopted in the current study. It was conducted during the 2015-2016 academic year in spring semester. Freshmen of Classroom Teaching Department comprised the universe of the study. The study group consisted of 99 students of Classroom Teaching Department at the Faculty of Education in Pamukkale University who attended the chemistry course of first-year curriculum.

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 115-121*

"Attitudes Towards Chemistry Course Scale" developed by Hançer, Uludağ and Yılmaz (2007) was used as data collection tool.

The participants were 80 female and 19 male students. 47 of the students graduated from Anatolian High School, 21 of them from regular high school, 25 of them from Teacher Training High School, and 6 of them from Vocational High School. 9 of the participants indicated that they had received the chemistry class before, 90 of the participants indicated that they had never received the chemistry class before. The Cronbach Alpha internal consistency coefficient was calculated as 0,939 for the current study. In order to determine whether the data showed normal distribution, Kolmogorov-Smirnov analysis was conducted and the results showed that the data was normally distributed (Z=0,676; p>0,05).

**Findings**

In the current study, statistical analyses were conducted in order to investigate the attitudes of pre-service elementary teachers towards chemistry course according to their gender and high school type that they graduated from.

1. With the intent to answer the sub question of "Is there a relationship between the attitudes of pre-service elementary teachers towards the chemistry course and their gender?" independent t-test (independent samples) analysis was performed. There were no meaningful differences found between female and male students' attitudes toward chemistry (t97=0,964, p>0,05). The mean values were found to be 90,91 for female students and 85,58 for male students.

2. With the intent to answer the sub question of "Is there a relationship between the attitudes of pre-service elementary teachers towards the chemistry course and the high school type that they graduated from?" One-Way ANOVA was conducted. There were no meaningful differences found among the types of high schools from which the students graduated (F(3,95)=2,482, p>0,05).

**Discussion, Conclusion and Suggestions**

As for the findings of the current study that aimed to determine the attitudes of pre-service elementary teachers towards the chemistry course and whether there were any differences in term of gender and high school type that they graduated from, the following conclusions were made:

The independent sample t-test conducted to answer the first sub-question revealed no significant differences (t97=0,964, p>0,05). The means acquired indicated that neither the female nor male students have positive attitudes towards the chemistry course.

The One-Way ANOVA conducted to answer the second sub-question revealed no significant differences among the high school types that the students graduated from (F(3,95)=2,482, p>0,05). In this sense, the type of high school that the students graduated from seems not to have an effect on the attitudes of the students towards chemistry course.

The mean of the students' attitudes toward chemistry course was found as 89,89. This score indicates that, in general, students do not have a positive attitude towards the chemistry class. Besides, 90 of the students stated that they had not taken the chemistry course before but 9 of them had. These 90 students also indicated that they had attended the fields of Turkish-language and Math classes in high school and the last time they took chemistry course was when they were at the $9^{th}$ grade. Therefore, it can be concluded that attending the fields of Turkish-language and Math class in high school can be one of the reason why pre-service elementary teachers have negative attitudes towards chemistry course.

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 122-133*

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

*"Research Article"*

# Ten Years Emotional Intelligence Scale (TYEIS): Its Development, Validity and Reliability

**Kerem Coskun,[*,1] Yucel Oksuz[2], H. Bayram Yilmaz[2]**

[1]Department of Primary Education, Artvin Coruh University, Artvin, Turkey
[2]Department of Educational Sciences, Ondokuz Mayis University, Samsun, Turkey

***Abstract***

This study aims to develop a reliable and valid measurement instrument of emotional intelligence based on mixed model. Mixed model of emotional intelligence and literature on it were investigated, and then an item pool with 53 items was developed. 14 expert of emotional intelligence examined 53 items. In order to make the expert's judgments standardized, Lawshe Content Validity Ratio was used. As a result of the ratio analysis, 18 items were discarded from initial draft of the scale. Data were collected from 492 children for the exploratory factor analysis (EFA). EFA results indicated the scale includes unidimensionality. Confirmatory Factor Analysis (CFA) yielded good model fit indices. Results indicated that the scale is reliable and valid instrument in measuring emotional intelligence.

***Article Info***

***Received***
*11 November 2016*

***Accepted***
*12 February 2017*

***Key words***
Emotional intelligence, measurement and assessment of the emotional intelligence, scale development, exploratory factor analysis, confirmatory factor analysis

## 1. INTRODUCTION

Happier, more productive and peaceful way of life has become main agenda for all individuals. It is emphasized in the literature that IQ is not strong enough to predict success in life. Moreover, it is known that those who have higher level of social and emotional skills are happier, more successful in life.

Emotional intelligence (EI) has offered new paradigm for educationalists that try to explain success and adjustment to environment. Concept of the EI first was developed by Mayer and Salovey (1990). However Goleman (1995) made it popularized and publicized. Large body of the research has proved that EI has positive impact on educational attainment, social adjustment, happiness, and academic self-efficacy (Hen and Goroshit, 2012; Hogan, Parker, Wiener, Watters, Wood, & Oke, 2010; MacCann, Fogarty, Zeidner, Roberts, 2011; Mavrovelli and Ruiz, 2010; Newsome Day, & Catano, 2000; Qualter Gardner, Pope, Hutchinson, Whiteley, 2012; Tariq, Qualter, Roberts, Appleby, Barnes, 2013; Saklofske, Auistin, Mastoras, Beaton, & Osborne, 2012; Sanchez-Ruiz, Mavrovelli, Poullis, 2013; Van Der Zee, Thijs, & Schakal, 2002). However there are disagreements and conflicts about definitions, qualities, and

[*]Corresponding Author Email: keremcoskun@artvin.edu.tr

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 122-133*

conceptualization of the EI. Those disagreements have stemmed from measurement paradigm of the EI (Zeidner, Matthews, & Roberts, 2009).

There are mainly three streams in EI: ability model, mixed models, and trait model (Zeidner et al., 2009). Salovey and Mayer (1990) developers of the ability model, described as that EI is the capacity to recognize and manage emotions in ourselves and in others, process emotional information. In the ability model, EI is assumed as capability of carrying out accurate emotional reasoning (Mayer, Roberts, & Barsade, 2008). The ability model constructs emotion and reasoning under same phenomena. The model consists of four abilities (those accurately perceiving emotion, using emotion to facilitate thought, understanding emotion, and managing emotion) (Salovey and Mayer, 1990; Mayer, Salovey, Caruso, & Sitarenios, 2003; Mayer, Salovey, & Caruso, 2004). In the ability model, there is a close interaction among the skills. For instance a child cannot be efficacious without perceiving emotion in herself (Mayer and Salovey, 1997).

Mixed models, another approach to the EI, view the EI as an integration of skills and qualities such as personality and motivational dispositions that are necessary to use the EI in real life. Proponents of the EI (Goleman, 1998; Bar-On, 2006; Petrides, 2001; Petrides, Pita, & Kokkinaki, 2007) deal with a wide range of skills and competencies rather than to define it as a single construct. In other words, EI is explained through broad definitions such as noncognitive capability, competency, skill or emotionally intelligent behavior, and dispositions of personality (Bar-On, 2006; Boyatzis, Goleman, & Rhee, 2000; Petrides, 2001; Petrides and Furnham, 2003). Bar-On (2000) describes the EI as cluster of noncognitive skills that are necessary to cope with effectively environmental demands. Bar-On (2006) suggests that the EI is one of the main determinants of effective human behavior. Bar-On (1997) developed EI model consisting of intrapersonal capacity, interpersonal skills, adaptability, stress management, motivation, and general mood. The Bar-On model claims that the EI is a joint of interrelated competencies, skills, and facilitators that influence how effectively an individual understands and expresses himself, recognize emotions in others, has good relationships with others, and fulfill social and environmental pressures (Bar-On, 2006). Goleman (1998) model is another model in the mixed models. It has five sub-dimensions as self-awareness, self-management, empathy, motivation and social skills.

Trait model developed by Petrides (2001) is another approach to the EI. Trait EI is a constellation of self-perception of the lower level of personality constructs. Trait EI includes 15 facets as adaptability, low impulsiveness, self-esteem, self-motivation, stress management, trait happiness, trait optimism, assertiveness, relationship skills, social competence, trait empathy, emotional expression, emotional management, emotional perception, and emotional regulation (Petrides, 2001; Petrides, 2010).

The difference between the EI models stems from way of measurement and assessment of the EI (Mayer, Salovey, Caruso, 2008; Perez, Petrides, & Furnham, 2005; Wigelsworth Humphrey, Kalambouka, & Lendrum, 2010, Zeidner et.al., 2009). The ability model deals with measurement and assessment of the EI in the same way as traditional intelligence standard test measures and assesses. The ability model measures and assesses through performance-based test because of the fact that the ability model deals with the EI as a single construct and standard intelligence type. According to the ability model, the EI is the capacity in reasoning with emotions. Therefore, the EI can be measured and quantified through the way in which standard traditional intelligence is measured. Participants' response on the EI related tasks are measured and assessed in accordance with such objectively right answer that measurement and assessment of the EI capabilities through the ability model does not include any bias or exaggerated evaluation of emotional capabilities. However, measurement and assessment in the

ability model are tough, not easy to administer due to the fact that expert panelists are needed to assess which respond is true, make decision about what respond is right according to objective rules (Wigelsworth et al., 2010; Wilhelm, 2005).

There are several instruments aiming to measure the EI related skills through the ability model and performance based tasks. Salovey and Mayer (1990) developed four branch of the EI, and devised the Multi Factor Emotional Intelligence Scale (MEIS). However, it was not found satisfactory in terms of validity and reliability. Mayer et al. (2002) developed the Mayer Salovey Caruso Emotional Intelligence Test (MSCEIT) to attenuate lengthy MEIS and ameliorate psychometric properties of the MEIS. Construct validation of the MSCEIT via confirmatory factor analysis by Rossen, Kranzler, & Algina (2008) revealed that the MSCEIT does not cover all constructs developed by Mayer et al. (2002), although Mayer, et al. (2003) founded that the MSCEIT has good model fit indices.

Furthermore, Fan, Jackson, Tang, & Zhang (2010) suggested that three factor solution of the MSCEIT has the best fitting model. Mayer et al. (in press) designed the MSCEIT Youth Version for children and youth between the ages 10 and 18 years. Peters, Kranzler, & Rossen (2009) investigated the MSCEIT-YV's construct validity and criterion-related related validity and concluded that it is a valid instrument in measuring emotional intelligence based on the ability model. Similarly, Rivers, Brackett, Reyes, Mayer, Caruso, & Salovey (2012) found that the MCEIT-YV produces valid results in measuring emotional intelligence among children aged from 10 to 13.

Emotional Intelligence Scale for Children (EISC) was developed by Sullivan (1999) through the ability model. However, internal consistency between subscales of the EISC varied low to moderate. Freudenthaler and Neubauer Emotional Intelligence Performance Test is another instrument use to assess emotional intelligence through performance-based approach and the ability model in EI (Freudenthaler and Neubauer, 2003). Emotional Accuracy Research Scale was developed by Mayer and Geher (1996) in accordance with the ability model. Both of the scales do not have any child or adolescent form.

The mixed models make emotions quantifiable through self-report. Self-assessment of emotions assumes that participants are competent enough to evaluate how much they have quality in emotions or their behaviors about the EI skills. In contrast to the ability model and performance based assessment, self-report of emotional responds may not have any objective criteria. Therefore, it is easy to administer and evaluate. However, this kind of assessment of emotions is risky. Participants may have such bias towards their EI skills that they can overrate their emotional intelligence skills. In order to reduce this risk, responds of participants through self-report can be checked with different source of information. For instance, responds of children can be compared and checked with observation checklist of teachers and evaluation of parents (Perez et al., 2005; Wigelsworth et al., 2010; Wilhelm, 2005; Zeidner et al., 2009).

There are numerous scales measuring the EI via self-report. Emotional Quotient Inventory developed by (Bar-On, 1997) is a self-report inventory with 133 items. Bar-On and Parker (2000) devised its youth version that measures the EI of children adolescents who are aged between 7 and 18 years. Another seminal measurement instrument of the EI is Trait Emotional Intelligence Questionnaire (TEIQue) developed by Petrides (2001). Petrides et al.(2006) adapted it to child and adolescent characteristics by shortening its length and named as Trait Emotional Intelligence Questionnaire- Adolescent Short Form (TEIQue-ASF). The TEIQue-ASF consists of 30 items, two for each of the 15 facets of Trait Emotional Intelligence and measures global trait EI. Its internal consistency reliability coefficient was found as 0.84. In addition to that, Cooper and Petrides (2010) tested its psychometric construction by using item-response theory and found that TEIQue-ASF has good psychometric properties. However,

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 122-133*

the fact that the TEIQue and TEIQue-ASF consist of too broad definitions and sub-dimensions, has drawn considerable criticism (Wigelsworth et al., 2010).

### 1.2. Purpose of the research

There are self-report emotional intelligence scales but they do not have any child form (Dulewicz and Higgs, 2001; Gignac, 2010; Palmer and Stough, 2002; Schutte Malouff, Hall, Haggerty, Cooper, Golden, & Dornheim, 1998; Tapia, 2001; Tett, Fox, Wang, 2005; Van Der Zee et al., 2002).

In this present study, an emotional intelligence scale, which measures emotional intelligence through self-report and are originated from Goleman (1998) conceptualization. There are two essential reasons why the TYEIS was developed for the children who are 10 years old. The first reason is about requirements of measurement of emotional intelligence through self-report. Measuring emotional intelligence via self-report assumes that participants in the sample have an insight about their social and emotional skill in depth and are objective, consistent, and genuine in assessing those skills. Age of 10 is a period in which metacognitive awareness, abstract reasoning, and objective thinking without being impressed with events, and objects begin to emerge among children. Therefore, they can be efficacious in assessing emotional skills through self-report in themselves. When developmental characteristics of primary school children are taken into consideration, 10 years old primary school children are more competent and efficacious to assess and evaluate emotional intelligence skills more accurately than younger children.

The second reason is about gender characteristics. Gender differences are clear between early childhood and age of 8 in favor of female children with respect to emotional intelligence skills. However, this difference disappears between 10 to 12 years because of more increase in male children's emotional intelligences (Keefer et al., 2013). Therefore, during primary school process, age of 10 is a period in which both female and male children are equal in terms of emotional intelligence skills.

When the literature is closely investigated, it can be seen that emotional intelligence scales for children and adolescents were designed in accordance with the Ability Model, the Bar-on Model, the Trait Emotional Intelligence Model but there is no emotional intelligence scale which originated from Goleman's conceptualization of the EI. Therefore, existing scale were grounded on such different models were there is no use in modifying them. Therefore, the present study aims to develop valid and reliable instrument of the EI based on Goleman's conceptualization of the EI.

## 2. METHODOLOGY

The aim of the present study is to develop a self-report emotional intelligence for primary school children so as to measure and assess level of social emotional learning, and reveals its psychometric properties. Item development, content validity, structural validity, reliability, and validity analysis were orderly carried out in the development process. The present study consists of two factor analysis as Exploratory Factor Analysis (EFA) discovering factor structures, internal consistency coefficients and Confirmatory Factor Analysis (CFA) which investigates how well data fit into previously revealed factor structures (DeVellis, 2012).

### 2.1. Participants

791 primary school children studying four grade and aged ten years old participated the study from different regions of Turkey in order to ensure representation of the sampling. Sample of exploratory factor analysis consists of 492 children, as sample of the confirmatory factor analysis includes 399 children.

## 2.2. Process

Studies of Goleman (1995, 1998) were scrutinized to build theoretical framework for the items. Moreover, several studies about the EI and its models were investigated in depth (Bar-On, 2006; Boyatzis et al., 2000; Humphrey et al., 2007; Killick, 2006; Mayer and Salovey, 1995; Mayer et al., 2004; Mayer et al., 2008; Petrides and Furnham, 2000; Perez et al., 2005; Warwick and Nettlebeck, 2004; Wigelsworth et al., 2010; Zeidner et al., 2009). On the other hand, 23 fourth grader children were asked to write a composition describing good and bad persons whom they encounter in their daily living so as to write appropriate items for 10 years old children and closely comprehend their emotional and social characteristics. Initially, 53 items were prepared in accordance with the literature review and compositions from the children. After constituting of the item pool, 53 items were formatted for expert investigation by inserting them into three points grade as *'Essential'*, *'Useful but not essential'*, and *'not necessary'*. The Content Validity Ratio developed by Lawshe (1975) was employed to make expert feedback standardized and ensure systematic content validity. Therefore, an expert panel was composed and comprised 14 experts whose expertise is on the EI. The Content Validity Ratio was determined as 0.51 for 14 expert panelists (Lawshe, 1975). After feedback from the expert panelists was received, 18 items were decided to remove from draft of the scale. Draft of the scale for the EFA was formed by placing 35 items onto three points scale as *'not true'*, *'somewhat true'* and, *'completely true'*.

## 2.3. Item analysis

Before the EFA, item analysis was conducted according to the corrected item total correlation. The corrected item-total correlation coefficient discovers the items that does not correlate the scale overall and measure different dispositions or characteristics and obstruct constructs. It was decided that the items whose item-total correlation coefficient is less than 0.30 discarded from the EFA. As a result of the item analysis, 5., 8.,9.,10., 11., 14., 16.,17., 19., 21., 22., 23., 24., 25., 28., 29., 30., 31., 33., 34. and 35 were excluded and 1., 2., 3.,4., 6., 7., 12., 13., 15., 18., 20., 26., 27. 32. were included in the EFA (Everitt, 2002; Field, 2009; Nunnally ve Bernstein, 1994). Initially those items' internal consistency coefficient was calculated and 1., 6., 15., and 26. Items were discarded from the EFA because of the fact that they caused a decrease in internal consistency coefficient. Consequently, based on the item analysis, the EFA was carried out with ten items.

**Table 1.** Results of Item Analysis

| Item No | Value of Corrected-Item Correlation | Item No | Value of Corrected-Item Correlation |
|---|---|---|---|
| Item 1 | .304 | Item 19 | .188 |
| Item 2 | .362 | Item 20 | .380 |
| Item 3 | .533 | Item 21 | -.012 |
| Item 4 | .427 | Item 22 | .186 |
| Item 5 | -.30 | Item 23 | .174 |
| Item 6 | .368 | Item 24 | .044 |
| Item 7 | .518 | Item 25 | .192 |
| Item 8 | .109 | Item 26 | .351 |
| Item 9 | -.198 | Item 27 | .460 |
| Item 10 | -.139 | Item 28 | .129 |
| Item 11 | -.407 | Item 29 | .151 |
| Item 12 | .407 | Item 30 | .181 |
| Item 13 | .427 | Item 31 | .165 |

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 122-133*

***Exploratory Factor Analysis (EFA):*** The EFA is a statistical process that enables one to identify inter-correlated variables and cluster them under same constructs (Field, 2009; Harrington, 2008; Rummel, 1967). In the EFA process, Kaiser-Meyer-Olkin (KMO) coefficient and Barlett Test are necessary to determine whether data is suitable for the EFA. KMO Coefficient was found as 0.93, and Barlett Test was significant ($X^2$*=2056, 806; p ≤ 0.001)*. These findings indicated that the sample is large enough to conduct the EFA (Field, 2009; Henson and Roberts, 2006; Pohlman, 2004; Thompson, 2004). Varimax rotation method makes factors such as interpretable clusters by maximizing dispersion of loadings that it was chosen as rotation method (Field, 2009). Eigenvalues were employed to make a decision about the number of factors. Eigenvalue indicated that there is one factor whose eigenvalue is more than 1. Therefore, it was decided that the scale includes one factor with 10 items (Field, 2009; Pohlman, 2004). It was also observed that one factor solution with 10 items explains 50% of total variance. According to Merenda (1997) number of factor must explain at least 50% of total variance. Consequently, this value was found as enough for identifying strong construct from the data. Factor loadings of the items in the one factor solution ranged between 0.433 and 0.818. As for reliability, overall internal consistency coefficient of the scale was found to be 0.89.

**Table 2.** Exploratory Factor Analysis Results

| Item | Factor Loadings | M | SD | Alpha If Item Deleted |
|---|---|---|---|---|
| Item 18 | .818 | 2.27 | .91 | .86 |
| Item 3 | .811 | 2,26 | .90 | .86 |
| Item 4 | .785 | 2,27 | .93 | .87 |
| Item 32 | .778 | 2,27 | .87 | .87 |
| Item 27 | .725 | 2,26 | .91 | .87 |
| Item 13 | .713 | 2,21 | .92 | .87 |
| Item 7 | .711 | 2,24 | .83 | .87 |
| Item 12 | .595 | 2,22 | .88 | .88 |
| Item 2 | .514 | 2,56 | .70 | .89 |
| Item 20 | .433 | 2,14 | .82 | .89 |
| *Eigenvalues* = 4,98 | | *Total Variance Explained*: 50% | | *KMO* =.93 |
| *Barlett Test*: $X^2$= 2056, 806; p ≤0.001 | | | | |

*M: Mean, SD: Standard Deviation*

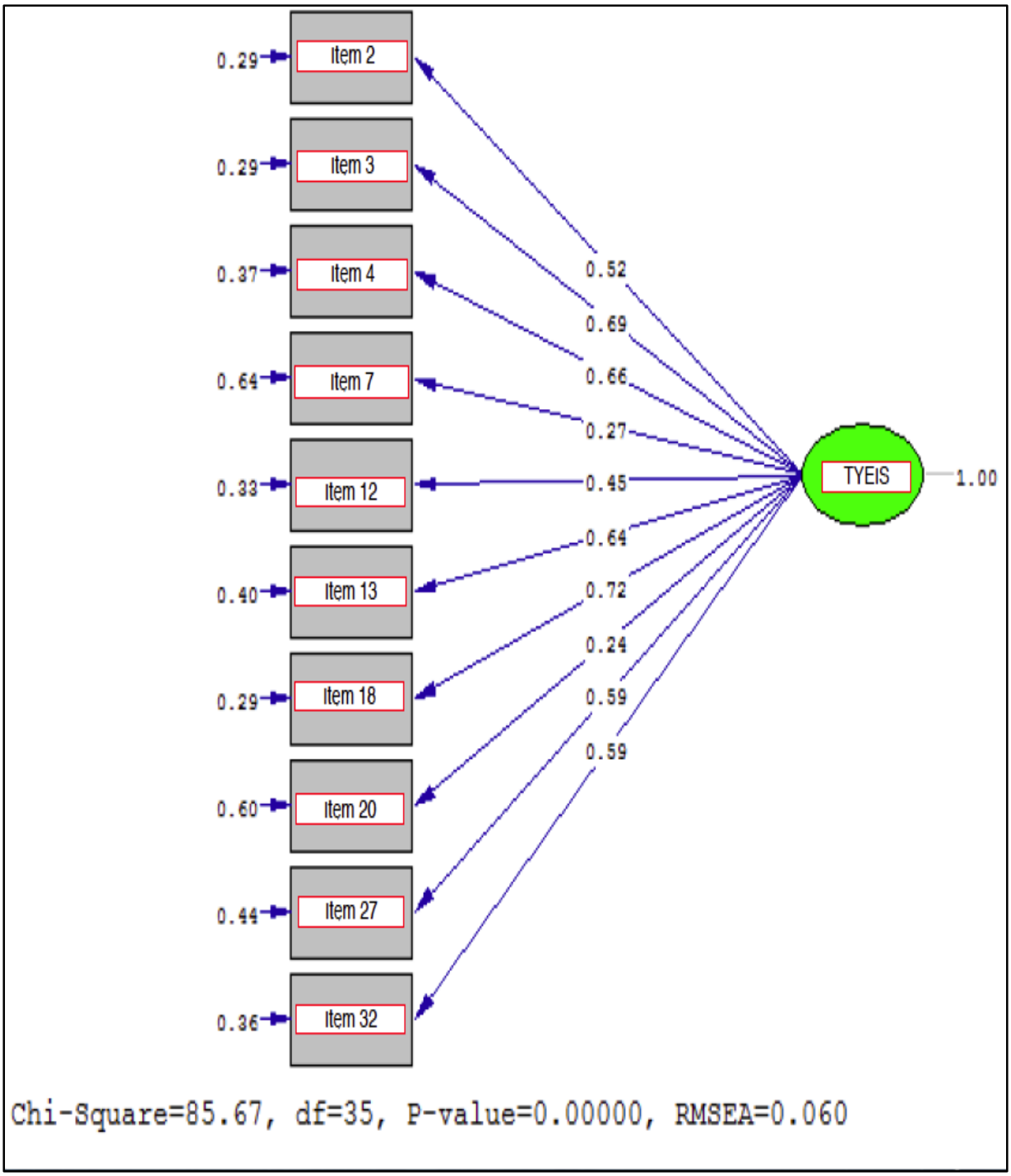
Based on findings about the EFA, single factor solution is reliable construct to measure the EI through self-report. It was decided that the Scale was named as Ten Years Emotional Intelligence Scale (TYEIS).

### 2.4. Confirmatory factor analysis (CFA)

The CFA is a factor analysis which reveals whether a defined model is confirmed or not, and previously determined factors are related to each other. Furthermore, the CFA determines such construct validity that the CFA allows researchers to accept or refuse the model**.** The CFA was conducted based on several fitting indices rather than single fitting index in order to test the model in depth (Harrington, 2008; Thompson, 2004). The TYEIS consisting of one factor and 10 items was applied on 399 children. In the CFA, results on $x^2$/df, RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Indices), IFI (Incremental Fit Index), GFI (Goodness of Fit Index), AGFI ( Adjusted Goodness of Fit Index), NFI (Normed Fit Index),

and RFI (Relative Fit Index) were reported. It was found that RMSEA is 0.06, CFI is 0.97, IFI is 0.9, RFI is 0.93, GFI is 0.95, AGFI is 0.94, NFI is .95, SRMR is 0.03. These findings indicate that the model with one factor has good fit indices.

**Figure 1.** Result of the Confirmatory Factor Analysis



*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 122-133*

## 3. DISCUSSION

This research reported the development and validation of the TYEIS which measures the EI through self-report. It consists of one factor and comprises 10 items. The TYEIS is based on Goleman (1998) conceptualization of the EI. On the other hand, it is a typical emotional intelligence scale for

The literature is full of instruments measuring social and emotional aspects of learning. However, their conceptualizations are originated from different concepts such as social and emotional skills, social competence, emotional competence and emotional literacy. The EI is one of the concepts about social and emotional aspects of learning (Wigelsworth et al., 2010). However, researches about the EI focus on adults while development of the EI for children is scarce (Peters et al., 2009).

EI scales for children are adaptation of adult scales to child characteristics. These scales are MSCEIT: Youth Version (Mayer et al., in press), EQI: Youth Version (Bar-On and Parker, 2000), TMMS-C (Rockhill and Greener, 1999), TEIQ: Adolescent Form. However, the TYEIS is typically developed for primary school children. Therefore, the TYEIS is confined to children who are at the age of 10.

There is difference in terms of conceptualization among those scales. The MSCEIT:YV was constructed upon Salovey and Mayer (1990) emotional intelligence model, the EQI:YV is based on Bar-On (1997) emotional intelligence model, the TEIQ: AV is framed within Trait Emotional Intelligence Model developed by Petrides (2001) while the TYEIS is based on Goleman (1995, 1998) conceptualization of the EI which is a mixed model.

The TYEIS is such a self-report emotional intelligence scale that it displays similarity with EQI:YV, TEIQue: ASF in terms of ways of measuring emotional intelligence. On the other hand, there is a difference between MSCEIT:YV and the TYEIS due to the fact that the MSCEIT:YV measures the EI through performance based approach.

## 4. CONCLUSION

The present study was conducted to develop the TYEIS, and confirm its reliability and validity through the EFA and the CFA. The item pool with 53 items was constituted through literature review on the EI, and compositions of the 23 children. The items were placed in three point grade such as *'Essential'*, *'Useful but not essential'*, and *'Not necessary'* to prepare for expert review. In order to ensure standardization in expert review, the Content Validity Ration was used. For this reason, an expert panel consisting of 14 experts was composed.

The Content Validity Ratio was determined as 0.51 due to the number of experts (Lawshe, 1975). As a result of the Content Validity Ratio Results, 18 items were removed from final form before the EFA. 492 primary school children, who are 10 years old, attended the EFA. Before the EFA, item analysis was carried out and 25 items were discarded from the EFA. Results of the KMO and Barlett Test indicated that the sample is large enough to conduct the EFA. There is single factor construct which account for 50% percent of total variance.

Overall, internal consistency coefficient was found as 0.89. After the EFA, the scale was named as Ten Years Emotional Intelligence Scale (TYEIS). The TYEIS with one factor and ten items was conducted on 399 children for the CFA. Results of the CFA revealed that the TYEIS with single factor solution has good model indices. Based the results, it was concluded that the TYEIS is a reliable and valid instrument in measuring and assessing the EI of primary school children through self-report.

The TYEIS can be used by teachers to evaluate impact of the activities on the EI and monitor students' emotional development. Besides, researchers can employ it to investigate correlation between the EI and other variables, to reveal impacts of the EI on various variables. Moreover, prospective studies whose purpose is to test its reliability and validity on children who are either younger or older than age of 10 can be carried out.

## 5. REFERENCES


Bar-On, R. (1997). *The Emotional Intelligence Inventory (EQ-I): Technical Manual*. TorontoMulti-Health Systems.

Bar-On, R. (2000). Emotional and social intelligence: Insights from the Emotional Quotent Inventory. In R. Bar-on & J.D.A. Parker (Eds.), *Handbook of emotional intelligence*, (pp: 363-388). San Francisco: Jossey-Bass.

Bar-On, R. (2006). The Bar-on model of emotional-social intelligence. *Psicothema*, 18, 13-25.

Bar-On, R., & Parker, J. D. (2000). *Bar-On Emotional quotient inventory (EQ-i-YV): Youth version*. Toronto: Multi-Health Systems.

Boyatzis, R. E., Goleman, D., & Rhee, K. (2000). Clustering competence in emotional intelligence: Insights from the Emotional Competence Inventory (ECI). In R. Bar-on, J.D.A. Parker (Eds.), *Handbook of emotional intelligence* (pp: 343-362). San Francisco: Jossey-Bass.

Cooper, A., & Petrides, K. V. (2010). A psychometric analysis of the Trait Emotional Intelligence Questionnaire–Short Form (TEIQue–SF) using item response theory. *Journal of Personality Assessment*, *92*(5), 449-457.

DeVellis, R.F. (2012). *Scale development*. California: Sage Publications.

Dulewicz, S.V., & Higgs, M.J. (2001) *EI general and general 360 user guide.* Windsor: NFER-Nelson.

Dulewicz, V., & Higgs, M. (1999). Can emotional intelligence be measured and developed. *Leadership & Organization Development Journal*, *20*(5), 242-253.

Everitt, B. S. (2002). *The Cambridge dictionary of statistics*. Cambridge: Cambridge University Press.

Fan, H., Jackson, T., Yang, X., Tang, W., & Zhang, J. (2010). The factor structure of the Mayer–Salovey–Caruso Emotional Intelligence Test V 2.0 (MSCEIT): A meta-analytic structural equation modeling approach. *Personality and Individual Differences*, *48*(7), 781-785.

Field, A. (2009). *Discovering statistics using SPSS*. London: Sage Publications.

Freudenthaler, H.H., & Neubauer, A.C. (2003) *The localization of emotional intelligence within human abilities and personality.* Poster presented at the 11th Biennial Meeting of the International Society for the Study of the Individual Differences (ISSID), Graz, Austria.

Gignac, G. (2010). Seven-factor model of emotional intelligence as measured by Genos EI. *European Journal of Psychological Assessment*, *26*(4), 309-316.

Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. New York: Bantam Books.

Goleman, D. (1998). *Working with emotional intelligence*. New York: Bantam Books.

Harrington, D. (2008). *Confirmatory factor analysis*. New York: Oxford University Press.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393-416.

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 122-133*

Hogan, M. J., Parker, J. D., Wiener, J., Watters, C., Wood, L. M., & Oke, A. (2010). Academic success in adolescence: Relationships among verbal IQ, social support and emotional intelligence. *Australian Journal of Psychology*, *62*(1), 30-41.

Humphrey, N., Curran, A., Morris, E., Farrell, P., & Woods, K. (2007). Emotional intelligence and education: A critical review. *Educational Psychology*, *27*(2), 235-254.

Keefer, K. V., Holden, R. R., & Parker, J. D. (2013). Longitudinal assessment of trait emotional intelligence: Measurement invariance and construct continuity from late childhood to adolescence. *Psychological Assessment*, *25*(4), 1255-1272.

Killick, S. (2006). *Emotional literacy at the heart of the school ethos*. London: SAGE.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563-575.

MacCann, C., Fogarty, G. J., Zeidner, M., & Roberts, R. D. (2011). Coping mediates the relationship between emotional intelligence (EI) and academic achievement. *Contemporary Educational Psychology*, *36* (1), 60-70.

Mavroveli, S., & Sánchez-Ruiz, M. J. (2011). Trait emotional intelligence influences on academic achievement and school behavior. *British Journal of Educational Psychology*, *81*(1), 112-134.

Mayer, J. D. & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. Sluyter (Eds.). *Emotional development and emotional intelligence: Implications for educators* (pp. 3-31). New York: Basic Books.

Mayer, J. D., & Salovey, P. (1995). Emotional intelligence and the construction and regulation of feelings. *Applied and Preventive Psychology*, *4*(3), 197-208.

Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology*, *59*, 507-536.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2002). *Mayer– Salovey–Caruso Emotional Intelligence Test (MSCEIT) item booklet.* Toronto: MHS Publishers.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2004). Target Articles: Emotional intelligence: Theory, findings, and implications. *Psychological inquiry*, 15(3), 197-215.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2008). Emotional intelligence: new ability or eclectic traits?. *American Psychologist*, *63*(6), 503.

Mayer, J. D., Salovey, P., & Caruso, D. R. (in press). *Mayer– Salovey–Caruso Emotional Intelligence Test: Youth Version-Research Edition (MSCEIT-YV).* Toronto: MHS Publishers.

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2. 0. *Emotion*, *3*(1), 97-105.

Mayer, J.D., & Geher, G. (1996). Emotional intelligence and the identification of emotion. *Intelligence*, 2, 89-113.

Merenda, P. F. (1997). A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement and Evaluation in Counseling and Development*, 30, 156-154.

Newsome, S., Day, A. L., & Catano, V. M. (2000). Assessing the predictive validity of emotional intelligence. *Personality and Individual Differences*, *29*(6), 1005-1016.

Nunnally, J. C. Bernstein. I. H. (1994). *Psychometric theory*. New York: McGraw Hill.

Palmer, B.R., & Stough, C. (2002) *Swinburne University Emotional Intelligence Test (Workplace SUEIT). Interim technical manual (Version 2).* Victoria: Swinburne University of Technology.

Pérez, J. C., Petrides, K. V., & Furnham, A. (2005). Measuring trait emotional intelligence. In R. Schulze, R.D. Roberts (Eds.). *Emotional intelligence: An international handbook*, (pp: 181-201). Massachusetts: Hogrefe Publishing.

Peters, C., Kranzler, J. H., & Rossen, E. (2009). Validity of the Mayer—Salovey—Caruso Emotional Intelligence Test: Youth Version—Research Edition. *Canadian Journal of School Psychology*, *24*(1), 76-81.

Petrides, K. (2001). *A psychometric investigation into the construct of emotional intelligence* (Unpublished Doctoral Dissertation). University College London, London.

Petrides, K. V. (2010). Trait emotional intelligence theory. *Industrial and Organizational Psychology*, *3*(2), 136-139

Petrides, K. V., & Furnham, A. (2003). Trait emotional intelligence: Behavioral validation in two studies of emotion recognition and reactivity to mood induction. *European journal of Personality*, *17*(1), 39-57.

Petrides, K. V., Pita, R., & Kokkinaki, F. (2007). The location of trait emotional intelligence in personality factor space. *British Journal of Psychology*, *98*(2), 273-289.

Petrides, K. V., Sangareau, Y., Furnham, A., & Frederickson, N. (2006). Trait emotional intelligence and children's peer relations at school. *Social Development*, *15*(3), 537-547.

Pohlmann, J. T. (2004). Use and interpretation of factor analysis in The Journal of Educational Research: 1992-2002. *The Journal of Educational Research*, *98*(1), 14-23.

Qualter, P., Gardner, K. J., Pope, D. J., Hutchinson, J. M., & Whiteley, H. E. (2012). Ability emotional intelligence, trait emotional intelligence, and academic success in British secondary schools: A 5year longitudinal study. *Learning and Individual Differences*, *22*(1), 83-91.

Rieffe, C., Terwogt, M. M., Petrides, K. V., Cowan, R., Miers, A. C., & Tolland, A. (2007). Psychometric properties of the Emotion Awareness Questionnaire for children. *Personality and Individual Differences*, *43*(1), 95-105.

Rivers, S. E., Brackett, M. A., Reyes, M. R., Mayer, J. D., Caruso, D. R., & Salovey, P. (2012). Measuring emotional intelligence in early adolescence with the MSCEIT-YV psychometric properties and relationship with academic performance and psychosocial functioning. *Journal of Psychoeducational Assessment*, *30*(4), 344-366.

Rossen, E., Kranzler, J. H., & Algina, J. (2008). Confirmatory factor analysis of the Mayer–Salovey–Caruso emotional intelligence test V 2.0 (MSCEIT). *Personality and Individual Differences*, *44*(5), 1258-1269.

Rummel, R. J. (1967). Understanding factor analysis. *Journal of Conflict Resolution*, 11(4), 444-480.

Saklofske, D. H., Austin, E. J., Mastoras, S. M., Beaton, L., & Osborne, S. E. (2012). Relationships of personality, affect, emotional intelligence and coping with student stress and academic success: Different patterns of association for stress and success. *Learning and Individual Differences*, *22*(2), 251-257.

Salovey, P., Mayer, J.D. (1990) Emotional intelligence. *Imagination, Cognition, and Personality*, 9 (3), 185-211.

Sanchez-Ruiz, M. J., Mavroveli, S., & Poullis, J. (2013). Trait emotional intelligence and its links to university performance: An examination. *Personality and Individual Differences*, *54*(5), 658-662.

*International Journal of Assessment Tools in Education: Vol. 4, Issue 2, (2017) pp. 122-133*

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., & Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. *Personality and individual differences*, *25*(2), 167-177.

Sullivan, A. K. (1999). *The emotional intelligence scale for children* (Unpublished Doctoral Dissertation). University of Virginia, Virginia.

Tapia, M. (2001) Measuring emotional intelligence. *Psychological Reports,* 88, 353-365.

Tariq, V. N., Qualter, P., Roberts, S., Appleby, Y., & Barnes, L. (2013). Mathematical literacy in undergraduates: role of gender, emotional intelligence and emotional self-efficacy. *International Journal of Mathematical Education in Science and Technology*, *44*(8), 1143-1159.

Tett, R. P., Fox, K. E., & Wang, A. (2005). Development and validation of a self-report measure of emotional intelligence as a multidimensional trait domain. *Personality and Social Psychology Bulletin*, *31*(7), 859-888.

Van der Zee, K., Thijs, M., & Schakel, L. (2002). The relationship of emotional intelligence with academic intelligence and the Big Five. *European Journal of Personality*, *16*(2), 103-125.

Warwick, J., & Nettelbeck, T. (2004). Emotional intelligence is…?. *Personality and Individual Differences, 37*(5), 1091-1100.

Wigelsworth, M., Humphrey, N., Kalambouka, A., & Lendrum, A. (2010). A review of key issues in the measurement of children's social and emotional skills. *Educational Psychology in Practice, 26*(2), 173-186.

Wilhelm, O. (2005). Measures of emotional intelligence: Practice and standards. In R. Schulze, R.D. Roberts (Eds.), *Emotional intelligence: An international handbook*, (pp: 131-154). Massachusetts: Hogrefe Publishing.

Zeidner, M., Matthews, G., & Roberts, R. D. (2009). *What we know about emotional intelligence: How it affects learning, work, relationships, and our mental health*. Massachusetts: MIT press.

*Int. J. Asst. Tools in Educ., Vol. 4, Issue 2, (2017) pp. 134-148*

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

*"Review Article"*

# Assessing Metacognition: Theory and Practices

**Nesrin OZTURK**[1]

[1]Ege University, Izmir, Turkey

***Abstract***

Many researchers in education emphasized students' metacognition should be fostered for academic development and achievement. However, to support students' metacognitive development and adequacy appropriately, their metacognition is to be assessed first. For this purpose, this theoretical study conducted a short review of metacognition, its assessment, and limitations of assessment measures and procedures. By focusing on ten current studies, a pattern of metacognition assessment was portrayed. It was concluded that knowledge about and regulation of cognition was assessed simultaneously as metacognition theory proposes. To assess especially knowledge about cognition, exclusively off-line measures were used. For regulation of cognition, both off-line and on-line measures were used. Chronological analysis of these studies revealed that latest metacognition assessment studies tended to utilize domain-specific or real-life tasks. Based on the findings, research implications for assessment and instruction were laid down.

***Article Info***

***Received***
*23 February 2017*

***Revised***
*12 March 2017*

***Accepted***
*17 March 2017*

***Keywords***

Metacognition,
Assessing metacognition,
Off-line measures,
On-line procedures

## 1. Introduction

Educating metacognitive individuals is one of the primary objectives of today's major initiatives since in 21st century, students should be able to build strong content knowledge by responding to varying demands of audiences, tasks, purposes, and disciplines by critically synthesizing different resources and valuing sound evidence. However, without metacognitive assessment that can provide with diagnostic information and directions for its instruction, educational initiatives seem to take students' metacognitive development or adequacy unreliably for granted.

[1] Corresponding Author Email: ozturknesrin@gmail.com

2148-7456 /© 2017 DOI: 10.21449/ijate.298299

This paper, thereby, focuses on assessing metacognition to contribute to its instruction. For this purpose, conceptual definitions and a short review of metacognition theory will be presented initially to disseminate the focus of assessment. Next, common procedures and measures used to assess metacognition and some limitations will be presented because they may confound interpretations. Then, recent research studies on assessing metacognition will be reviewed analytically to detect whether and how metacognition theory is exercised for assessment purposes. Finally, possible future research and implications for metacognition assessment and instruction will be discussed.

### 1.1. Conceptual Definitions

The definition of metacognition has important implications for its assessment considering the construct validity. The common conceptualization of metacognition pertains to knowledge about cognition and regulation of cognition (Flavell, 1979). To create the framework for this paper and for the studies to be selected, this paper adopted Block's definition of metacognition assessment. According to Block (2006), metacognitive assessment pertains to assessing "reader's awareness and knowledge of the mental processes engaged during reading… [and] if a reader can monitor, regulate, and direct their thoughts before, during, and after reading to obtain a complete comprehension of text" (p. 84). Expanding this definition on learning in general, this paper defines metacognition assessment as assessing individuals' knowledge about and regulation of cognitions (planning for the task, monitoring one's performance, regulating skills, and evaluating performance and goal fulfilment). In the following, the fundamentals of these definitions will be elaborated.

## 2. Metacognition and Components of Metacognition

Jacobs and Paris (cited in Michalsky, Mevarech, & Haibi, 2009) described metacognition as "the conscious self-awareness of one's own knowledge of task, topic, and thinking, and the conscious self-management (executive control) of the related cognitive process" (p. 364). Almost 30 years later, Veenman, Van Hout-Wolters, and Afflerbach (2006) defined metacognition as "a higher-order agent overlooking and governing the cognitive system, while simultaneously being part of it" (p.5). Veenman et al. (2006) argued that if metacognition is a set of self-instructions to regulate task-performance, then cognition is the vehicle for these self-instructions. In order to understand this two-way mental processing and to conceptualize metacognition better, Nelson's (1996) Metacognitive Model of consciousness and cognition can be studied. Nelson (1996) distinguished "object-level" (cognitions concerning external objects) and "meta-level" (cognitions concerning cognitions of external objects) processes and by his Metacognitive Model, it was highlighted that "any lower-level cognition can itself be the subject of a higher-level cognition" (Nelson, 1996, p. 105). That is,

> [i]nformation about the state of the object-level is conveyed to the meta-level through monitoring processes, while instructions from the meta-level are transmitted to the object-level through control processes. Thus, if errors occur on the object-level, monitoring processes will give notice of it to the meta-level and control processes will be activated to resolve the problem (Veenman et al., 2006, p. 4).

To understand these definitions and conceptualizations better, components of metacognition needs dissemination. *Knowledge about cognition* pertains to thinking and sensitivity to act accordingly (Flavell, 1979). It includes "students' declarative, procedural, and conditional

knowledge about cognition, cognitive strategies, and task variables that influence cognition" (Pintrich et al., 2000, p. 45). Declarative knowledge pertains to one's awareness of *what* influences cognitions and includes person, task, and strategy variables (Veenman et al., 2006). Procedural knowledge pertains to a large variety of strategies or skills (Pintrich, Wolters, & Baxter, 2000; Pressley, Borkowski, & Schneider, 1987; Veenman et al., 2006) and it reflects "an appreciation for *how* skills operate or are applied" (Cross & Paris, 1988, p. 131). On the other hand, conditional knowledge pertains to one's knowing *when* and *why* to use declarative and procedural knowledge (Garner, 1990).

Metacognition also includes *regulation of cognition.* It is generally categorized into three: planning, monitoring and regulation, and evaluation (Ozturk, 2016; Schraw, 1998). Planning pertains to goal-setting that guides cognitions in general and monitoring specifically (Pintrich et al., 2000). Although it is not easy to separate monitoring and regulating from each other during a task performance, these activities can be distinguished conceptually as in the following (Pintrich et al., 2000). Monitoring activities include assessing learning and performance-in-action while regulation pertains to changing cognitions and behaviour to match them with personal goals and task demands (Pintrich et al., 2000). Evaluation, lastly, pertains to "appraising the products and efficiency of one's learning" by re-visiting one's goals and conclusion (Schraw, 1998, p.115). However, although these facets are described separately here, it is important to recognize that knowledge about and regulation of cognition relate and have an interactive nature (Veenman et al., 2006).

## 3. Assessing Metacognition

In literature, metacognition is assessed by different procedures and measures. In the following, common measures and procedures will be disseminated with regards to metacognition components.

***Knowledge about cognition:*** Measures assessing knowledge of cognition can look similar to standard tests because knowledge of cognition is considered much like knowledge stored in memory (Pintrich et al., 2000). That is, individuals tell whether they know or do something or not. Baker and Cerro (2000) identified interviews and/or questionnaires as one of the most frequently used methods to assess metacognitive knowledge. Metacognitive Awareness of Reading Strategies Inventory (MARSI), for example, was developed to assess domain specific metacognition. Mokhtari and Reichard (2002) designed MARSI to assess adolescent and adult readers' metacognitive awareness and perceived use of reading strategies; global reading strategies, problem-solving strategies, and practical support strategies. On the contrary, Metacognitive Assessment Inventory (MAI), developed by Schraw and Dennison (1994), is used to measure adults' general metacognitive knowledge and regulation of cognition. These instruments are examples of off-line measures as they can be administered effectively to large groups and scored easily.

***Regulation of cognition*:** To measure metacognitive judgements, monitoring, and regulation, on-line processes are used. By these measures, individual are asked what they do and think before, during, and after a cognitive task. Procedures such as "detection of errors in passages; ratings of felt understanding; self-corrections during oral reading; completion of cloze tasks; on-line measures of processing during reading (e.g. eye movements and reading times); and

retrospective or concurrent verbal reports (e.g. thinking aloud)" can be used to assess individuals' regulation of cognition (Baker & Cerro, 2000, p.102).

To measure metacognitive monitoring, self-report judgments can be used (Pintrich et al., 2000). Before individuals perform some tasks, they can be asked to rank how easy the information will be to learn. Then, after given some tasks and study trials, individuals can be required to rank and make a judgment of their learning. Because individuals' confidence in their performance is assessed by comparing it to their actual performance, the accuracy of their judgements relates to their monitoring ability (Pintrich et al. 2000). That is to say, students who felt they know something and did, and students who felt they did not know something and did not are both considered good monitors as they can make accurate judgements.

Regulation can be assessed by several different questionnaires and interview protocols such as the Learning and Study Strategies Inventory (LASSI), the Motivated Strategies for Learning Questionnaire (MSLQ), the Self-Regulated Learning Interview Schedule (SRLIS). The MSLQ and LASSI ask individuals to respond to Likert-type items for their domain- general and domain-specific cognitive strategy use and regulation of cognition, respectively. The MSLQ is designed to assess rehearsal, elaboration, organization, and critical thinking while metacognitive monitoring and self-regulation are assessed on a 12-item scale apart from resource management strategies (Pintrich et al., 2000). Moreover, the SRLIS asks individuals about self-regulation considering specific tasks. After individuals are presented some descriptions of the content, they are asked how they would behave during a) a classroom discussion, b) short writing assignment, c) mathematics assignment, d) end-of-term test, e) homework assignment, and f) studying at home (Zimmerman & Martinez-Pons, cited in Pintrich et al., 2000). The responses are categorized into knowledge, monitoring behaviour, strategy use, and regulation. Similarly, Survey of Reading Strategies (SORS), developed by Sheorey and Mokhtari (2001), intends to measure the perceived use of strategies while reading academic materials. On a 5 point Likert-scale, individuals are asked to indicate the frequency of reading strategy use.

Veenman (2005) categorized measures for cognitive regulation into three as prospective, concurrent, and retrospective measures. Collected before a learning task, prospective measures aim at identifying metacognitive skills either in general or prior to specific learning tasks. Veenman (2005) stated that questionnaires can be used for this purpose and individuals can be asked to indicate to what extent and/or how often a statement represents their study behavior on for example, a Likert-scale. Apart from questionnaires, Veenman (2005) also appreciated interview techniques as a form of prospective measures. By structured or hypothetical interview procedures, individuals can be assessed for their strategy usage. While their answers are coded, the number of the strategies and metacognitive merit can be evaluated (Veenman, 2005).

Concurrent measures help collect data during individuals' task performance. A predominant method for assessing metacognitive skills is the analysis of think aloud protocols (Veenman, 2005). The basic principle of think aloud is that "participants are instructed to merely verbalize their thoughts during task performance. Only in case they fall silent, the assessor may urge them to "keep on talking" (p.80). Think aloud protocols can specifically be utilized for assessing individuals' monitoring of the text characteristics, understanding, problems in comprehension, and their strategic processes used to comprehend text (Pressley & Afflerbach, 1995). Think aloud processes are transcribed verbatim and analysed according to a coding scheme, resorting exclusively to the quantity of metacognitive activities and the quality of metacognitive processes. The protocols are generally analysed by two or more judges separately for inter-rater reliability.

In relation to evaluation and judgements of metacognitive activities, Veenman (2005) especially warned the assessors not to confuse correctness of knowledge to mindfulness.

Veenman (2005) also highlighted systematical observations can be used to assess metacognitive skills. The observations are made by the judges who are physically but unobtrusively present during task performance. Judges can also watch videotapes afterwards to score individuals' metacognitive behaviours if there are concerns related to their presence within the site. Often used with young children, on-line observations can only account for quantitative behavioural assessment, not for the metacognitive objectives. As in the case of think aloud, a coding scheme should describe all possible metacognitive activities to be evaluated.

The error detection paradigm is another approach to assess metacognitive skills (Baker & Cerro, 2000). Individuals are presented with texts that contain problems and/or errors and their metacognitive ability is inferred from their attention to the embedded errors. The underlying assumption of this paradigm is that these problems or errors disrupt comprehension and the readers who monitor their comprehension notice them. Baker and Cerro (2000) stated whether readers are capable of detecting the errors can be assessed by performance measures such as underlying errors, verbal reports during reading, and on-line measures like eye-tracking.

Retrospective measures, on the other hand, are administrated just after a performance has been completed. Due to the risk of memory failure and distortions, stimulated-recall technique that requires participants to review a video of their own performance can be used to help individuals with the reproduction of their though processes during their task-performances (Veenman, 2005).

## 4. Limitations of Current Assessment Approaches

Assessing metacognition is important but simultaneously it is challenging (Schraw, 2000). Despite numerous measures and procedures developed to meet this assessment challenge, metacognition that is a multi-layered complex phenomenon may not be easily assessed. While measures of metacognitive knowledge do not tap into metacognitive monitoring or regulation, metacognitive judgements and monitoring measures are not consistent in assessing the same components (Pintrich et al., 2000). Furthermore, regulation is commonly assessed rather than monitoring (Pintrich et al., 2000; Pressley & Afflerbach, 1995).

With regards to previously mentioned procedures and measures, some limitations will be discussed in the following. One of the frequently used methods, verbal reports possess some limitations which should not be ignored for accurate interpretations. During the interviews, it is possible that individuals do not understand the questions and do not ask for clarifications, or they may not be willing to express their genuine thoughts and experiences (Baker & Cerro, 2000). Their responses, therefore, might be indecisive and socially desirable ones. Moreover, as Veenman (2005) argued, it is never for sure whether the respondents have metacognitive strategies and skills at their disposal or they can really use them when appropriate even though they can report the relevance. Also, as Pintrich et al. (2000) stated that although participants can be asked for a number of strategies during the interviews, they may not include domain-specific control and regulation strategies. In addition to these limitations, some concerns with interpretation cannot be ignored. As Whitebread and colleagues (2009) emphasized, interpreting self-reports and scoring especially open-ended questions is not an easy task. Such a task requires not only expertise in data analysis, but it also requires expertise with metacognition theory and its practical applications.

Regarding questionnaires, Veenman (2005) stated that although they are relatively easy to administer, questionnaires do not reliably describe metacognitive behaviour. Reviewing 21 questionnaire studies, Veenman and van Hout-Wolters (as cited in Veenman, 2005) also stated that the predictive value is low; the mean variance accounted in learning outcomes was around 3%. Students' individual reference points may cause this low predictive value because students might compare themselves with the best or poorest classmates. Moreover, as Veenman (2005) and Pintrich et al. (2000) stated, measuring and evaluating skills through questionnaires is a very controversial issue. Not only can questionnaire items portray individuals' adequacy with regulation of cognition, but also the representativeness of such questionnaires might be problematic regarding the limited number of items on questionnaires. For these reasons, reliability, construct and structural validity, mismatch between theoretical models of metacognition and subcomponents requires careful interpretations. Moreover, generalizability of these measures might be problematic considering diverse students characteristics (Pintrich et al., 2000).

Furthermore, there are limitations with think-aloud protocols. While think-aloud aims to understand metacognitive and cognitive processes, it is important to remember that these processes cannot be always accessible to consciousness. Individuals may not be always aware of their knowledge, monitoring, or regulation or their verbal proficiency might not be adequate to describe these. Think aloud may also slow down or interrupt cognitive processing and might limit some individuals' working memory capacity (Baker & Cerro, 2000; Lai, 2011; Veenman, 2005). Although all these factors can be controlled well enough, still personal and/or affective factors (such as motivation, anxiety, self-esteem, verbal ability, age, expertise, and individuals' knowledge) might interfere with individuals cognitive processing (Baker & Cerro, 2000; Pintrich et al., 2000; Schraw & Moshman, 1995). Therefore, there is a risk that interpreting think-aloud procedures might underestimate metacognitive capacity (Lai, 2011). To recognize confounds in disguise, think-aloud protocols should be scored by judges with sufficient expertise and experience with metacognition theory.

Furthermore, in spite of providing some evidence for on-line comprehension monitoring, error detection paradigm has limitations. First of all, Baker and Cerro (2000) emphasized that depending on readers' being informed about the problems in the text, differences in their comprehension monitoring can occur. Also, reliance on verbal-reports, as mentioned beforehand, might not always be trustworthy. In addition, as readers might use variety of criteria for detecting errors and evaluating their understanding, problems that individuals report might be completely different than those intended to be conveyed. However, failure to notice particular problems in a text does not necessarily portray poor comprehension. Moreover, error detection paradigm is also criticized for ecological validity; individuals do not normally read texts embedded with errors. Although individuals' monitoring strategies can be assessed by the error detection paradigm, it is not for certain whether these individuals monitor their comprehension under normal conditions without any stimuli like texts used for error detection.

Systematic observations, which are somewhat independent of confounds like individuals' verbal ability and working memory capability, still have limitations. Considered to be more ecologically valid compared to the previous paradigms (Lai, 2011), observations need to be converged with other measures for construct validity (Veenman, 2005). This is because it cannot assess metacognitive intentions for performing certain behaviours (Veenman, 2005). Although systematic observations are considered to take social processes of learning into consideration and embedded in the context of instruction, the judgment is limited to the observants' inferences. Even

the construct of metacognition is standardized and checklists are developed, because of social influence and other contextual factors, the inferences derived from metacognitive assessment might not be always accurate.

Lastly, stimulated-recall technique holds drawbacks in assessing metacognitive skills. This is basically due to the time lag between individuals' actual performances and their verbal reports. When participants watch their own performances, it might be difficult for them to reproduce memory traces and covert mental activities. Therefore, instead of correct recollections, reconstructive interpretations may be elicited (Veenman, 2005). As Nisbett and Wilson (cited in Veenman, 2005) stated, even retrospective verbal reports of higher order processes might lack accuracy because participants might tell more than they know.

The limitations of particular approaches covered in this theoretical study pertain to individuals' working memory capacity, verbal proficiency, personal performance criteria, tendencies towards socially desirable responses, observant' expertise and interpretation biases, and measures', procedures', and interpretations generalizability. Therefore, one needs to make informed choices about the measures and procedures to serve the purposes, needs, and the context best (Pintrich et al., 2000).

## 5. Research on Metacognition Assessment

In this part, ten research studies whose focus is assessing metacognition in the domain of reading will be presented. To understand how metacognition theory and previous research on metacognition impact current assessment practices, these studies will be analysed for their definition of metacognition, assessment measures and procedures, and their limitations, if stated at all. Also, selected studies will be presented chronologically to recognize whether there is an emerging pattern in the assessment of metacognition while its literature keeps increasing.

Kolić-Vehovec and Bajšanski (2006) aimed to explore students' developmental differences ($5^{th}$ to $8^{th}$ grade) in comprehension monitoring and perceived use of reading strategies. For this purpose, they used error correction and text sensitivity tasks from Metacomprehension test. Although it is difficult to separate monitoring from regulation, their study was built on the argument that comprehension monitoring is important for the regulation of reading and regulation is manifested in a way how readers plan, monitor, evaluate, and use available information while they are building comprehension. Besides, because "the ability to monitor their [readers'] comprehension is not enough guarantee that children actually use reading strategies" (p.441), a self-report measure of reading strategies use was also adopted. While the results revealed significant grade level differences for text comprehension and cloze task performances, there were no statistically significant differences for error detection and text sensitivity among grade levels. Besides, comprehension monitoring was found to be significantly correlated to reading comprehension. However, perceived use of reading strategies was correlated to reading comprehension only in eighth grade.

Desoete (2008) also assessed third-graders' metacognitive skillfulness. For this purpose, she investigated four skills; prediction, planning, monitoring, and evaluation and calibration by using the Prospective Assessment of Children (PAC), Retrospective Assessment of Children (RAC), and teacher ratings as off-line ratings, and think-aloud protocol. Moreover, EPA 2000 was used as a combined (prospective and retrospective) form of assessment. The results confirmed teacher

ratings on predictions skills positively correlate with the combined assessment measure, but not with the child questionnaire. Teacher ratings of evaluation skills also correlated with the concurrent and combined assessment techniques. Besides, overall teacher ratings correlated with prospective child measure. Children's prospective and retrospective questionnaire results, which was not much influenced by students' actual performance, were not different and showed some evidence for convergent validity. The evaluation skill was found to be relatively independent in prospective child ratings and think-aloud. The author also highlighted "high intercorrelations between prediction, planning, monitoring, and evaluation skills rated by the teachers and between the prediction and evaluation skills assessed by EPA2000" (p. 204). Think aloud protocols, on the other hand, showed some evidence for the interaction of monitoring, planning, and prediction skills. Although the skills are generally related, the author recommended assessing skills separately.

Aiming to investigate Turkish high school students' metacognition and its relation to achievement goals, Sungur and Senler (2009) examined students' metacognition by its preliminary components. For this purpose, the study utilized the Metacognitive Awareness Inventory (MAI), the Achievement Goal Questionnaire (AGQ), the Competence Expectancy Scale, and the Challenge and Threat Construals. After running a confirmatory factor analysis, the authors pointed out that participants had "reasonable knowledge about themselves as learners, about strategies, and when and how to use these strategies. They also appeared to regulate their cognition at high levels" (p.52). It was also stated that all types of goal orientation and knowledge and regulation of cognition were positively correlated at each level.

Turan, Demirel, and Sayek (2009) argued that metacognitive awareness and self-regulated learning skills are important especially in the field of medicine because of the rapid change in knowledge. Conducting their study at four different medical schools implementing different curriculum, the authors used self-regulated learning perception scale (SRLPS) and metacognitive awareness inventory (MAI) to collect data from 862 students. They found a statıstically significant difference among medical school curricular models. MAI and SRLSP scores of the students who study a problem-based learning (PBL) curriculum were higher than discipline- and system- based curricular models.

Zhang (2009), acknowledging the importance of reading in a second language, pointed out that non-native readers can apply their native language knowledge of reading processes and strategies to second and/or foreign language contexts. For effective strategy instruction, the study aimed to assess students' metacognitive awareness and reading strategy use and examine whether there are any differences in strategy choice among different proficiency levels. For these purposes, the author used SORS. The analysis revealed that the participants use reading strategies at a high-frequency level; they showed a moderate to high usage with problem solving strategies as their primary choice, followed by global strategies and support strategies. However, high-, intermediate-, and low-proficiency students were different in their strategy choice; "their pattern of strategy use is closely related to their overall EFL achievement" (p. 48).

Onovughe and Hannah (2011) also examined secondary school students' awareness and utilization of metacognitive strategies to comprehend academic materials. To obtain data from a group of 120 students, the authors used a questionnaire called "Students' Awareness and Application of some strategies to Reading and Comprehension" (p.344). While students' awareness of reading skills and strategies were rated on a 2-point scale, a set of 5 questions was used to identify students' purposes for reading. The authors concluded that secondary school

students in their study were aware of metacognitive strategies to a large extend as over 60% affirmation was obtained for each aspect of metacognitive strategies. Moreover, these participants applied metacognitive strategies in reading and comprehension to a large extent. The authors also highlighted a correlation between metacognitive awareness and utilization of metacognitive strategies.

Lee, Teo, and Bergin (2009) conducted their study specifically to understand "whether regulation of cognition and knowledge of cognition are related to everyday problem solving and whether students who perform better in the decision-making problem will better differentiate the various components of metacognition" (p. 89). The authors recruited 254 fifth grade students and they were given an everyday decision-making type of problem to solve; how to select a bike for purchase. To understand children's decision-making, the authors adapted MAI for the problem-solving scenario. The findings revealed that 30.6% of the variance was accounted for regulation of and knowledge about cognition. And "at the higher level of decision-making, knowledge of cognition and regulation of cognition were differentiated in their use by the participants" (p. 97). The authors, therefore, claimed that students, who made poorer decisions in the given problem, could not discriminate among components of metacognition.

Akyol and Garrison (2011) examined how students demonstrate their metacognitive knowledge and skills in an online learning context. Coding 16 undergraduate students' responses for knowledge about cognition, monitoring, and regulation of cognition, the authors chose 3 weeks ($1^{st}$, $5^{th}$, and $9^{th}$) of online discussions to assess students' metacognition. Observing possible changes in metacognition over time, the authors stated that while knowledge of cognition decreased in time, monitoring and regulation of cognition was noted to increase over time.

The study carried out by Saraç and Karakelle (2012) investigated the interrelation between different on-line and off-line measures for assessing metacognition. Working with 47 fifth grade elementary students, the authors utilized teacher rating scale, self-report questionnaire (Jr. MAI), think aloud protocols, and accuracy ratings (JOL) of text comprehension. The results showed some evidence for the correlation between two off-line measures (positive) and online measures (negative). However, there was no significant correlation between off-line and on-line measures.

Arguing that metacognitive skills directly shape learning behaviour and consequently impact learning outcomes, Veenman, Bavelaar, De Wolf, and Van Haaren (2014) conducted a study to assess metacognitive skills. As they argued that metacognitive skills can be assessed by on-line measures, students' log-files of computerized tasks were used as data sources. Still, because log-files cannot reflect their metacognitive consideration for the specific enactments, log-file analysis was validated against other on-line methods. 52 students performed a computerized inductive learning task and then they were asked to complete a performance post-tests. The results revealed high convergent validity between log-file indicators and human judgements of learner activities.

## 6. Critical Summary

This analysis of ten recent studies confirmed that knowledge about and regulation of cognition was assessed simultaneously in most cases as metacognition theory presents them. In this review, eight studies exclusively used off-line measures to assess metacognition (see Table 1). By using questionnaires, these eight studies assessed metacognition although questionnaires have been criticized especially for not appropriately assessing metacognitive skills. While two studies

used both online and offline measures, only one study used solely online measures to assess metacognition. Also, only two studies focused solely on regulation of cognition rather than integrating it with knowledge about cognition. One of these studies assessed regulation of cognition through online measures and the other utilized both online and offline measures.

While research has used different measures and procedures to assess metacognition, in this review, a total of eight studies used different off-line measures like MAI, MARSI, Jr. MAI, and SORS to assess knowledge about cognition. Only one study used an on-line measure of assessing knowledge about cognition. In that study, metacognitive behaviours were recorded and inferences regarding participants' knowledge about cognition were made by the researchers. Regulation of cognition was assessed in all studies. In addition to aforementioned measures, different self-report measures and on-line measures were used to assess regulation of cognition. However, only five studies assessed regulation of cognition by on-line measures like error correction and text sensitivity, think-aloud, observation of metacognitive behaviours, and analysis of computerized tasks' log-files. Besides, despite not mentioned in the literature, two of the studies used teacher-ratings to validate students' metacognition.

Few studies declared limitations that stem from their measurement choices. Although previous studies and pioneers in the field explicitly pointed out the limitations of recent measurement approaches, most of the researchers in this review were concerned about sample size, participant characteristics, and/or contexts that they collected their data from, if they ever mentioned limitations. Considering the generalizability of their findings and replicating similar research, one needs to be cautious of and alert against the potential flaws of the measurement, as well.

Lastly, the chronological analysis of these studies enabled to detect an emerging pattern in assessing metacognition. The latest studies in this review included specific tasks to assess metacognition rather than assessing it as a rigid construct. The earlier studies tended to use domain-general off-line measures to assess metacognition. The latest studies, on the contrary, included more specific real-life tasks for which participants need to employ different cognitive skills. While participants were engaged in task completion, their metacognition was assessed through on-line measures. Instead of generalizing one's metacognitive capability, such assessment procedures shed light on metacognitive processes and capabilities at the moment.

**Table 1.** Metacognition assessment pattern

| Components | Type | Off-line Assessment | On-line Assessment | Extras | Total |
|---|---|---|---|---|---|
| *Knowledge of Cognition* | Declarative<br>Procedural<br>Conditional | ✓ MAI<br>(Metacognitive Awareness Inventory)<br>✓ MARSI<br>(Metacognitive Awareness of Reading Strategies Inventory)<br>✓ Jr. MAI<br>✓ SORS<br>(Survey of Reading Strategies) | ✓ Metacognitive behaviours | | 8 |
| *Regulation of Cognition* | Predication<br>Planning<br>Monitoring<br>Regulation and Evaluation<br>Calibration | ✓ PAC and RAC<br>(Prospective Assessment of Children & Retrospective Assessment of Children)<br>✓ MAI<br>✓ Jr. MAI<br>✓ MARSI<br>✓ SORS<br>✓ JOL<br>(Judgment of Learning) | ✓ Error correction and text sensitivity<br>✓ Think-aloud<br>✓ Metacognitive behaviours<br>✓ Decision-making behaviours<br>✓ Log-files of computerized tasks | Teacher-ratings | 10 |

*Note*. Based on 10 studies assessing metacognition (published after 2006).

## 7. Discussion and Conclusion

Metacognition, a profound predictor of learning (Wang, Haertel, & Walberg, 1990), is composed of interacting features of knowledge about cognition and regulation of cognition (Schraw & Dennison, 1994). To assess metacognition, two approaches have been used. Through off-line methods, knowledge about strategies and estimated performance can be measured either before or after tasks. However, it cannot guarantee or estimate that individuals have strategies at their disposal or use them to regulate their learning behaviour (Veenman et al., 2006). Despite this fact, this review study confirmed that research unanimously used questionnaires to assess knowledge about and regulation of cognition. When people are given certain options to choose among, they are not really asked to manifest their knowledge, but they are asked to pick an appropriate option. Still more importantly, through questionnaires a researcher may not discriminate whether metacognitive knowledge is correct or complete or whether one can appreciate the usefulness of such knowledge in a situation. Moreover, interpretations of such assessment practices might be misleading when one might be inclined to generalize the assessment results, obtained by for example MAI, to any learning and/or performance situations. Nevertheless, I do not propose eliminating questionnaires to assess metacognition, but I propose integrating different data sources for verification.

Moreover, while assessing metacognition, one needs to recognize that interpretations are based on specific cases. Generalizing individuals' metacognitive adequacy to any other similar domains, therefore, might be inappropriate. Future research assessing individuals' metacognition can benefit from different domain tasks and cross-compare metacognitive engagement or behaviours to develop a holistic understanding of metacognitive adequacy. For example, to assess monitoring, instead of just asking students to detect errors in a reading paragraph, they may also be asked to reflect their understanding of a math problem, which has for example, logical inconsistencies. Then, individuals' metacognition in different domains can be analysed and compared.

Moreover while assessing metacognition, it is important to recognize different factors might impact metacognitive engagement and it is possible to confound these to students' adequacy. For example, when individuals are graded for their performances, as a partial fulfilment of their degrees, achievement motivation can interfere with the interpretations. On the other hand, individuals might not be interested in the task that they are provided and therefore, they may not be motivated for task completion. Without acknowledging characteristics and potential impacts of tasks and without recognizing individuals' volitional control (Pressley & Afflerbach, 1995), metacognitive assessment interpretations might be biased or incomplete. Future research on metacognitive assessment, therefore, needs to consider what drives or stops individuals from engaging in metacognitive processes and actions.

Finally, before assessing metacognition, it is very crucial to state the purpose of the assessment explicitly. While "research" and theory development can be valid reasons for academia, there should be some practical implications for teachers and students. As Lai (2011) stated, metacognition is not assessed regularly and traditionally at schools. Its instruction, therefore, might likely be ignored despite its beneficiary merits for achievement unless instructional and assessment practices are intertwined. While metacognition assessment research is carried out, it is important to state how metacognition assessment can benefit its instruction. In relation, as mentioned beforehand, two of the studies used teacher-ratings to validate individuals' self-reports of metacognition. Although metacognitive instruction has not been given a voice in these studies and teachers' awareness of metacognition and skills to

teach for metacognition has not been assessed, teachers' ratings were used to validate students' metacognition. Similar studies adopting teacher- ratings need to examine whether and how teachers interpret and rate students' metacognition especially in case they might not be metacognitive or they might not teach for metacognition, at all. In such cases, teachers, in fact, might know what and how to assess exactly and validly. Therefore, future research had better relate metacognition instruction and diagnostic assessment practices to empower not only students' metacognition but also teachers' understanding and practices of metacognition instruction and assessment.

## 8. References

Akyol, Z., & Garrison, D. R. (2011). Assessing metacognition in an online community of inquiry. Internet and Higher Education, 14(3), 183–190. https://doi.org/10.1016/j.iheduc.2011.01.005

Baker, L., & Cerro, L. (2000). Assessing metacognition in children and adults. In G. Schraw & J. C. Impara (Eds.), Issues in the measurement of metacognition (pp. 99–145). Lincoln, NE: Buros Institute of Mental Measurements.

Block, C. C. (2006). What are metacognitive assessments? In S. E. Israel, C. C. Block, K. L. Bauserman, & K. Kinnucan-Welsh (Eds.), Metacognition in literacy learning: Theory, assessment instruction, and professional development (pp. 83–100). New Jersey: Lawrence Erlbaum Associates, Inc.

Cross, D. R., & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology*, 80(2), 131–142. https://doi.org/10.1037/0022-0663.80.2.131

Desoete, A. (2008). Multi-method assessment of metacognitive skills in elementary school children: How you test is what you get. *Metacognition and Learning*, 3(3), 189–206. https://doi.org/10.1007/s11409-008-9026-0

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911.

Garner, R. (1990). When children and adults do not use learning strategies: Toward a theory of settings. *Review of Educational Research*, 60(4), 517–529. https://doi.org/10.3102/00346543060004517

Kolić-Vehovec, S., & Bajšanski, I. (2006). Metacognitive strategies and reading comprehension in elementary-school students. *European Journal of Psychology of Education*, 21(1983), 439–451. https://doi.org/10.1007/BF03173513

Lai, E. R. (2011). Metacognition : A Literature review research report. Research Reports. New York, NY:Pearson. Retrieved from http://www.datec.org.uk/CHAT/chatmeta1.htm

Lee, C. B., Teo, T., & Bergin, D. (2009). Children's use of metacognition in solving everyday problems: An initial study from an Asian context. *Australian Educational Researcher*, 36(3), 89–102. https://doi.org/10.1007/BF03216907

Michalsky, T., Mevarech, Z. R., & Haibi, L. (2009). Elementary school children reading scientific texts: Effects of metacognitive instruction. *The Journal of Educational Research*, 102(5), 363–376. https://doi.org/10.3200/JOER.102.5.363-376

Mokhtari, K., & Reichard, C. A. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology*, 94(2), 249–259.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, 51(2), 102–116. https://doi.org/10.1037/0003-066X.51.2.102

Onovughe, G., & Hannah, A. (2011). Assessing ESL students' awareness and application of metacognitive strategies in comprehending academic materials. *Journal of Emerging Trends in Educational Research and Policy Studies* (JETERAPS), 2(5), 343–346.

Ozturk, N. (2016). An analysis of pre-service elementary teachers' understanding of metacognition and pedagogies of metacognition. *Journal of Teacher Education and Educators*, 5(1), 47–68.

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In G. Schraw & J. C. Impara (Eds.), Assessing metacognition and self-regulated learning (pp. 43–97). Lincoln, NE: Buros Institute of Mental Measurements.

Pressley, M., & Afflerbach, P. (1995). Verbal protocols of reading: The nature of constructively responsive reading. NJ: Routledge.

Pressley, M., Borkowski, J. G., & Schneider, W. (1987). Cognitive strategies: Good strategy users coordinate metacognition and knowledge. In R. Vasta & G. Whitehurst (Eds.), Annals of Child Development, Vol. 5 (pp. 89–129). Greenwich: JAI Press.

Saraç, S., & Karakelle, S. (2012). On-line and off-line assessment of metacognition. *International Electronic Journal of Elementary Education*, 4(2), 301–315.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, 26(1), 113–125.

Schraw, G. (2000). Assessing metacognition: Implications of the Buros symposium. In G. Schraw & J. C. Impara (Eds.), Issues in the measurement of metacognition (pp. 297–321). Lincoln, Nebraska: Buros Institute of Mental Measurements.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, 19(4), 460–475. https://doi.org/10.1006/ceps.1994.1033

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, 7(4), 351–371. https://doi.org/10.1007/BF02212307

Sheorey, R., & Mokhtari, K. (2001). Differences in the metacognitive awareness of reading strategies among native and non-native readers. System, 29(4), 431–449. https://doi.org/10.1016/S0346-251X(01)00039-2

Sungur, S., & Senler, B. (2009). An analysis of Turkish high school students' metacognition and motivation. *Educational Research and Evaluation*, 15(March 2015), 45–62. https://doi.org/10.1080/13803610802591667

Turan, S., Demirel, O., & Sayek, I. (2009). Metacognitive awareness and self-regulated learning skills of medical students in different medical curricula. *Medical Teacher*, 31(10), e477–e483. https://doi.org/10.3109/01421590903193521

Veenman, M. V. J. (2005). The assessment of metacognitive skills. In B. Moschner & C. Artelt (Eds.), Lernstrategien und Metakognition: Implikationenfür Forschung und Praxis (pp. 75–97). Berlin: Waxmann.

Veenman, M. V. J., Bavelaar, L., De Wolf, L., & Van Haaren, M. G. P. (2014). The on-line assessment of metacognitive skills in a computerized learning environment. *Learning and Individual Differences*, 29, 123–130. https://doi.org/10.1016/j.lindif.2013.01.003

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, 1(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1990). What influences learning? A content analysis of review literature. *The Journal of Educational Research*, 84(1), 30–43.

Whitebread, D., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., Demetriou, D. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning*, 4(1), 63–85. https://doi.org/10.1007/s11409-008-9033-1

Zhang, L. J. (2009). Title Chinese senior high school EFL students' metacognitive awareness and reading-strategy use Chinese senior high school EFL students' metacognitive awareness and reading-strategy use. Source Reading in a Foreign Language, 21(1), 37-59. Retrieved from https://repository.nie.edu.sg/bitstream/10497/16307/1/RFL-21-1-37_a.pdf

*Int. J. Asst. Tools in Educ., Vol. 4, Issue 2, (2017) pp. 149-165*

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

*"Review Article"*

# A Generalizability Analysis of the Reliability of Measurements: "An Example of Cell Division and Heredity Unit"

**Gülşah BAŞOL**[*,1] **Muammer YÜKSEL**[2]

[1]Gaziosmanpaşa Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Böl., Taşlıçiftlik Yerleşkesi, 60100 Tokat, Türkiye.
[2]Ayşe Temizel Ortaokulu, 45500 Soma/Manisa, Türkiye.

***Abstract***

The purpose of the study is to measure students' performance through different measurement tools and compare the findings through G Theory in order to identify the errors associated with the raters and items to improve the future applications. The sample consisted of 48 eighth graders in Kars. Two different types of exams (a multiple choice test and an essay) were applied and essays were graded by three raters. G and K analyses were performed on the results. According to the findings, the error rate was higher for the essays in comparison to multiple-choice test. The mean score was higher for the multiple-choice test, the variances were similar. There were no differences among the essay scores given by different raters. Findings of decision study indicated Student facet as the main source of the variation in the data for both types of the measurements.

***Article Info***

***Received***
*03 February 2017*

***Revised***
*23 March 2017*

***Accepted***
*28 March 2017*

***Keywords***
Measurement,
Generalizability Theory,
Reliability,

# Bir Genellenebilirlik Analizi Çalışması: "Hücre Bölünmesi & Kalıtım Ünitesi"

**Özet**

Araştırmanın amacı, aynı konu alanı ile farklı ölçme araçlarından elde edilen puanların G Kuramı ile karşılaştırılmasıdır. Örneklem Kars'ta öğretim gören 48 sekizinci sınıf öğrencisinden oluşmuştur. İki farklı sınav türü (çoktan seçmeli test ve yazılı sınav) öğrencilere uygulanmış ve yazılı sınav üç puanlayıcı tarafından puanlanmış, sonuçlar üzerinde G ve K analizleri yapılmıştır. Sonuçlar çoktan seçmeli test ve yazılı sınav için karışan hata varyanslarının yazılı sınavda daha fazla olduğunu, çoktan seçmeli testin puan ortalamasının daha yüksek olduğunu, varyansların ise iki sınav türü için birbirine yakın olduğunu ortaya koymuştur. Yazılı sınav için puanlayıcılar arasında herhangi bir fark bulunmamıştır. Ayrıca, karar çalışmasından elde edilen sonuçlar verideki varyansın ana kaynağının Öğrenci faktörü olduğunu ortaya koymuştur.

***Makale Bilgisi***

***Makale Gönderim***
*03 Şubat 2017*

***Makale Düzeltme:***
*23 Mart 2017*

***Makale Kabul***
*28 Mart 2017*

***Anahtar Kelimeler:***
Ölçme,
Genellenebilirlik Kuramı,
Güvenirlik,

[*]Corresponding Authors' E-mails: gulsah.basol@gop.edu.tr yuksel.muammer.my@gmail.com

Bu çalışma, 1-3 Eylül 2016'da Antalya'da gerçekleştirilen V. Eğitimde Ölçme ve Değerlendirme Kongresi'nde bildiri olarak sunulmuştur.

2148-7456 /© 2017 DOI: 10.21449/ijate.303991

## 1. GİRİŞ

Son yıllarda ön plana çıkan yaşam boyu öğrenme anlayışı, bireylerin gelişmelerine büyük oranda katkı sağlamaktadır. Hayatımızın her safhasında yeni şeyler öğrenme ve kendimizi geliştirme imkanı sağlayan bu görüş, öğrenmelerin ve eğitimin önemini daha çok ortaya çıkarmaktadır. Eğitim süreci bünyesinde yer alan hazırbulunuşluk, güdülenme, öğretim, ölçme ve değerlendirme gibi kavramlar da artık hayatın içinde daha sık benimsenmektedir. Her biri birbirinin tamamlayıcısı ve önemli bir parçası olan içi içe geçmiş bu süreçleri bilmek eğitim ve öğretimin kalitesini arttırır.

Eğitim istendik davranış oluşturma veya istendik davranış değiştirme süreci olarak, toplumun süzgeçten geçirilmiş değerlerinin, ahlak standartlarının, bilgi ve beceri birikimlerinin yeni nesillere aktarılmasıdır (Senemoğlu, 2002). Eğitim süreci sonunda bireylerin belli konularda bilgi, beceri ve tutum kazanması beklenir. Bu istendik bilgi, beceri ve tutumların kazanılma düzeyinin; sürecin verimliliğini göstermesi ve dönüt sağlayarak süreci zenginleştirmesi beklenir. Eğitim sistemimizde bireylerin bu kazanımları başarı olarak nitelendirilmekte ve farklılaşan başarı düzeylerinin doğru olarak ölçülmesine çalışılmaktadır.

Etkinlikler sonunda beklenen kazanımların; bir kısmının oluştuğu, bir kısmının yeterli düzeyde oluşmadığı, istenmeyen kazanım şeklinde ortaya çıktığı veya planlandığı şekilde oluşmadığı görülmektedir. Bu durum eğitimde kontrol ihtiyacını doğurur (Turgut ve Baykul, 2010). Burada yer alan kontrol kavramı eğitim sürecinin ve ürünlerinin gözden geçirilmesi ve bir sonuca varılması anlamına gelmektedir. Kontrol süreci eğitimi hem planlı hale getirir hem de var olan eksikliklerin giderilmesine ve kalitenin arttırılmasına olanak sağlar.

Öğrencilerin başarılarının belirlenmesinde öncelikle ölçme ve sonrasında bunu da içine alan değerlendirme sürecine yer verilmelidir. Eğitim sürecindeki bireylerin eğitimden ne kadar yararlandıkları ya da öğrenilmesi beklenilen kazanımlara ne ölçüde ulaşıldığı sürekli merak konusudur. Çünkü hem eğitimin niteliği hem de bireyler hakkında verilecek kararlar için kazanımların ulaşılma düzeyleri saptanmalıdır. Burada da devreye ölçme ve değerlendirme süreci girer.

Kazanımla ifade edilen hedefleri gerçekleştirme yolunda öğretim etkinlikleri planlanır. Öğretimde izlenen yöntemi de dikkate alarak farklı ölçme araçları arasından, öğrenmenin gerçekleşip gerçekleşmediğini yoklamak için en uygun olanı seçilir. Değerlendirmenin amacına göre kullanılan ölçme araçları da çeşitlilik gösterir. Ölçme yönteminin hedeflenen kazanımlara uygun olması ölçme sonuçlarının geçerliği için önemlidir. Bu nedenle ölçülmek istenilen kazanımların niteliğine en uygun olabilecek ölçme aracının seçilmesine gerekli önem verilmelidir.

Kullanılacak ölçme aracına; öğrencinin hazırbulunuşluk düzeyi, sınavın yapılacağı ortam, zaman sınırlaması olup olmadığı ve uygulama koşulları gibi faktörler dikkate alınarak karar verilir. Farklı ölçme araçlarıyla elde edilen sonuçların benzer olup olmadığı araştırılmaya değer bir konudur. Bu sayede farklı kaynaklardan ulaşılan ölçme sonuçlarının güvenilir olup olmadığını anlamak mümkün olur. Eğitim sürecinin önemli bir parçası olan ölçme için bireylerin farklı ölçmeler neticesinde ortaya çıkan sonuçların birbirleri ile ilişkisinin nasıl olduğu bir merak konusudur. Bu soruların cevaplarının bulunması elbette performans, not ve başarı seviyesi olarak ilerleyen sürecin daha anlamlı şekilde açıklanmasını sağlayabilir.

Katılımcıların performansının ölçülmesinin amaçlandığı araştırmalarda araştırma sürecinde yer alan ve araştırmayı etkileyen ya da etkileyebilecek pek çok değişken kaynağı bulunmaktadır.

Bu değişken kaynaklarının etkilerinin olup olamadığı ya da etkilerinin ne ölçüde olduğunun ortaya konmasında farklı ölçme kuramlarından yararlanılmaktadır. Bu kuramlardan biri olan Genellenebilirlik Kuramı (G Kuramı); hata kaynaklarını aynı anda ele alması ve birbirleri ile ilişkilerine yer vermesi nedeni ile araştırmada değişken kaynaklarının birbirleri ile karşılaştırılmasına olanak vermektedir.

G Kuramı ölçme sonuçlarının güvenirliğinin belirlenmesini, güvenilir gözlemlerin tasarımını, araştırılmasını ve kavramsallaştırılmasını sağlayan istatistiksel bir kuramdır. G Kuramı, Klasik Test Kuramı (KTK)' nın bir uzantısıdır (Cronbach ve diğerleri, 1972; Brennan, 2001). G Kuramı, KTK' nın günümüzde hala popüler olan gerçek puan modelinin sınırlılıklarına olan cevap vermek amacıyla Cronbach ve arkadaşları (1963) tarafından ortaya atılmıştır. KTK, bir tek gerçek puana sahip her bir gözlem ya da test puanının paralel gözlemlerin bir grubuna ait tek bir güvenirlik katsayısı üretmesi fikri etrafında merkezlenir (Lord ve Novick, 1968; Baykul, 2000). G Kuramı ölçüm prosedürlerinin geliştirilmesine uygulanmış olmakla birlikte, özellikle eğitim araştırmaları içinde uygulaması sınırlı kalmıştır (Bottema-Beutel, Lloyd, Carter ve Asmus, 2014).

Shavelson ve Webb'e (1991) göre, G Kuramı dört farklı açıdan KTK' nın daha genişletilmiş bir halidir: 1. Genellenebilirlik Kuramı, çoklu varyans kaynaklarını tek bir analizde ele alır. 2. Her bir varyans kaynağının büyüklüğünün belirlenmesini sağlar. 3. Bireylerin performanslarına dayalı hem bağıl kararlar hem de mutlak kararlar alınmasına ilişkin iki farklı güvenirlik katsayısının (sırasıyla; G katsayısı ve Phi katsayısı) hesaplanmasına olanak sağlar. 4. Belirli bir amaca bağlı olarak, ölçme hatasının en aza indirgenebileceği ölçmelerin düzenlenmesine (Karar "K" çalışmaları) imkân tanır.

G Kuramı farklı hata kaynaklarının varyans analizi yoluyla ayrı ayrı ve bir arada rapor edilerek kestirmesini sağlar. Genellenebilirlik Kuramında yer alan çoklu hata kaynakları bir örnek üzerinden açıklanabilir. Bir başarı testinin iki ya da daha fazla puanlayıcı tarafından puanlandığı bir durumda, kestirilebilecek hata kaynağı ile aynı testin paralel formlarından elde edilen puanlara ilişkin kestirilen hata kaynağı aynı olmayacaktır. Klasik Test Kuramında bu hata kaynaklarını aynı anda kestirmek mümkün değildir (Güler, 2009).

G Kuramına göre değişkenlik kaynakları çapraz (crossed) ya da yuvalanmış (nested) şekilde olabilir (Rentz, 1987). Çaprazlanmış desende değişkenlik kaynağının koşulları başka bir değişkenlik kaynağının koşullarıyla örtüşmektedir (Brennan 2001). Çaprazlanmış desende değişkenlik kaynakları arasına 'x' işareti konulmaktadır. Araştırmada bir değişken kaynağı diğer değişken kaynağının tüm koşulları ile örtüşmüyor, sadece belli koşulları ile örtüşüyor ise bu çalışma desenine yuvalanmış desen denilmektedir. Yuvalanmış desende değişkenlik kaynakları arasına ' : ' işareti konulur.

G Kuramında güvenirliğin araştırılması iki aşamadan oluşmaktadır. Bunlardan ilki Genellenebilirlik çalışması (G-çalışması) ve ikincisi Karar çalışması (K-çalışması) şeklindedir (Kaya, 2011). G çalışması, ölçüm hatasını makul ve ekonomik olarak çok yönlü yalıtmak ve tahmin etmek, uygulama yapabilmek üzere tasarlanmıştır (Shavelson ve Webb, 2005). G çalışmasının amacı, ölçmenin birden çok kullanımını kestirmek ve bu sayede varyans kaynakları ile ilgili mümkün olan en çok bilgiye ulaşmaktır. G çalışması, mümkün olan en çok değişkenlik kaynağını içerecek biçimde tasarlanmalıdır. Bir başka deyişle G çalışması, kabul edilebilir gözlemlerin evrenini mümkün olan en geniş şekilde tanımlar (Shavelson ve Webb, 1991).

G-çalışması sürecinde, örneklemin evrene genellenebilmesi için, puanların değişkenliğinin tüm kaynakları (varyans bileşenleri) ve bunlar arasındaki etkileşimler aynı anda ANOVA yöntemi

kullanılarak kestirilmektedir. Kestirilen bu varyans bileşenleri bir sonraki aşama olan K-çalışmasında kullanılır (Kaya, 2011). G çalışması sonucunda elde edilen sonuçların K çalışmasında kullanımı söz konusudur ya da araştırmacı isterse devam etmeyip, çalışmasını G çalışması olarak sonlandırabilir.

K-çalışması, karar vermek üzere belirli bir amaç için veri toplanan çalışmadır ve yapılan bir K çalışmasında, incelenen bireyleri tanımlamak için veri toplanabilir (Kaya, 2011). Bir G çalışmasına karşılık, birden fazla K çalışması yapılabilir. K çalışması ile güvenirlik katsayısına benzeyen genellenebilirlik katsayısına (*G katsayısı*) ve güvenirlik indeksine (*Phi katsayısı*) ulaşılır. G katsayısı evren puanı varyansının kendisi ile bağıl puan varyansının toplamına oranıdır ve bağıl modellerde çalışılmaktadır (Çakıcı Eser, 2011).

G katsayısı KTK' daki güvenirlik katsayısına benzemektedir. G katsayısı, göreli karar modelinde gerçek varyansın, göreceli varyans ve gerçek varyansın toplamına bölünmesi ile bulunmaktadır. Öte yandan güvenilirlik endeksi ya da Phi (Φ) katsayısı mutlak karar modeli ile kullanılmaktadır. Phi katsayısı, gerçek varyansın, mutlak hata varyansı ve gerçek varyansın toplamına bölünmesi ile hesaplanır. Diğer bir deyişle, bu iki katsayı hatanın ne koşullarda kabul edileceğine göre farklılık göstermektedir (Alharby, 2006). Sonuç olarak, tek bir G-çalışmasından elde edilen aynı varyans kestirimlerine dayalı pek çok K-çalışması düzenlenebilir. K-çalışmasında kullanılan formül Spearman-Brown formülüne benzerdir (Mushquash ve O'Connor, 2006).

Cronbach ve arkadaşları tarafından 1963 yılında temelleri atılan G Kuramı ile ilgili çalışmalar yurt dışında aynı tarihleri takriben başlarken, ülkemizde 2004 yılından itibaren ve daha çok yüksek lisans ve doktora tezleri üzerinde yoğunluk göstermiştir. Bu yeni kuram; başlarda tezlerde yapılan araştırmalarla, günümüzde makalelerle ve üzerine yazılan bir kitap ile (Güler, Kaya Uyanık ve Taşdelen Teker, 2012) daha çok dikkat çekmeye başlamıştır.

Ülkemizde henüz yangınlaşmaya başlayan G Kuramı çalışmaları genellikle performansın ölçülme süreci, puanlayıcılar ve klasik ölçme araçları üzerinde yoğunlaşmıştır. Puanlayıcıların, bireylerin ve maddelerin etkileri araştırılırken farklı desenlerin incelendiği araştırmalar ( Wang, 2005; Au, Prahardhi ve Shiell 2008; Lane ve Sabers, 1989; Nalbantoğlu Yılmaz ve Uzun Başusta, 2012; Nalbantoğlu , 2009) daha çok yoğunluk kazanmıştır.

Atılgan (2004); Güler (2008) ve Alkahtani (2012) G Kuramı ile yaptıkları çalışmalarında, KTK yanında Çok Değişkenli Rasch Modelini (ÇDRM) kullanmışlar; maddelerin zorluk düzeyleri ve puanlayıcıların puan verme eğilimleri hakkında bilgiye ulaşmaya çalışmışlardır. Kuramların ve modellerin karşılaştırılmasının yanında, bazı çalışmalarda Lojistik Regresyon Analizi kullanılması, farklı kesme puanları hesaplama yöntemlerinin karşılaştırılması, farklı ölçeklerin güvenirliklerinin araştırılması çalışmaları G Kuramı yardımıyla yapılmıştır.

Araştırmada aşağıda yer alan alt problemlere cevaplar aranmıştır:

1. Çoktan seçmeli test için G Kuramına göre kestirilen parametrelerin varyansları ve toplam varyansları açıklama yüzdeleri nedir?

2. Çoktan seçmeli test için yapılan K çalışması sonuçlarına göre G ve Phi katsayılarının değişimleri nasıldır?

3. Yazılı sınav için yapılan G Kuramına göre kestirilen parametrelerin varyansları ve toplam varyansları açıklama yüzdeleri nedir?

4. Yazılı sınav için yapılan K çalışmasına göre farklı senaryolara göre G ve Phi katsayılarının değişimi nasıldır?

## 2. YÖNTEM

Başol (2008)'e göre betimsel araştırmalar ne ve nasıl sorularına sistematik olarak cevap vererek, olay ve durumların detaylı olarak betimlenmesi amacıyla yapılır. Araştırma, G Kuramı ile mevcut sistemde öğretmenlerin aynı konu hakkında kullandıkları ölçme araçları arasında ilişkiyi belirleme çalışması olduğundan betimsel bir araştırma niteliği taşımaktadır.

### 2.1. Evren ve örneklem

Araştırmanın çalışma evrenini 2013-2014 Eğitim-Öğretim yılında Kars il merkezinde öğrenim gören 8. sınıf öğrencileri oluşturmaktadır. Araştırma örneklemini ise Kars il merkezinde yer alan bir ortaokulda öğrenim gören 48 öğrenci oluşturmaktadır. Araştırmada uygulama kolaylığından dolayı amaçlı örnekleme gitmiştir.

### 2.2. Veri toplama aracı

Araştırma için gerekli veriler, araştırmacılar tarafından hazırlanan yazılı sınav (essay) ve ölçme sürecinde daha önce kullanılmış olan sorular arasından seçilen çoktan seçmeli test sorularına verilen cevaplardan elde edilmiştir**.** Araştırma soruları için 8. sınıf fen bilimleri dersinde yer alan 'Hücre Bölünmesi ve Kalıtım' ünitesine ait 20 kazanım ele alınmış olup öğrenci seviyesi de düşünülerek çoktan seçmeli test için ilk olarak 40 madde seçilmiştir. Hazırlanan bu sınav öncelikle iki konu alanı uzmanı ve bir dil uzmanına danışılarak deneme formatı için hazır hale getirilmiştir. Deneme uygulaması Kars il merkezinde yer alan farklı üç okulda öğrenim gören 96 öğrenci üzerinde yapılmış ve büyük ölçüde eksik olduğu belirlenen altı katılımcının cevapları çıkarılmıştır. Geriye kalan 90 kişinin cevapları dikkate alınmış ve deneme uygulamasının yapıldığı 90 kişiden oluşan grup nihai uygulamaya dahil edilmemiştir.

Deneme uygulaması için test ve madde istatistikleri TAP.exe (Brooks ve Johanson, 2003) uygulaması kullanılarak elde edilmiştir. Konu alanı ve kazanımların ağırlıkları incelenmiş ve alan uzmanların görüşüde alınarak 40 madde hazırlanmış ancak konu alanını daha iyi temsil ettiği ve bazı kazanımlar için yazılan soru sayısının dağılımın farklı olduğu için madde güçlük katsayıları ve madde ayırt edicilik güçleri incelenerek ağırlılıklı olarak orta güçlük seviyesinde, madde ayırtdiciliği .40 üzerinde olan maddeler seçilerek her biri dört seçenekli 22 maddelik coktan seçmeli testi oluşturmuştur.

Yazılı sınav için iki konu alanı uzmanının görüşüne başvurularak kapsam geçerliliğinin sağlanması amacıyla sorular hazırlanmış ve bir dil uzmanına danışılarak uygulama formu hazır hale getirilmiştir. Soruların yanlış anlaşılmalara neden olmaması ve tarafsızlığa hizmet etmesi açısından, bir kız ve bir erkek öğrenciye önceden çözdürülmüştür. Sınavın uygulandığı bu iki öğrenci için uygulanan sınav sonrası öğrenci görüşleri ele alındığında cinsiyete göre yanlılığının olmadığı sonucuna ulaşılmıştır. Ayrıca bu iki öğrenci nihai uygulama grubu arasında yer almamıştır.

Çoktan seçmeli test ve yazılı sınav Kars il merkezinde yer alan bir ortaokulda öğrenim gören 48 katılımcıya birer hafta ile uygulanmış ve uygulamalar birinci araştırmacı tarafından bireysel olarak gözlemlenmiştir.

### 2.3. Verilerin analizi

Araştırmacılar tarafından geliştirilen ölçme araçlarından elde edilen verilerin analizinde TAP.exe (Brooks ve Johanson, 2003) , SPSS (16. Sürüm, SPSS Inc, Chicago, 2007) ve G Kuramı

analizleri için EduG software (EduG version 6.1-e, Quebec, Canada, 2012) paket programları kullanılmıştır.

İlk olarak belirlenen ölçme araçları ile gerekli uygulamalar yapılmış, çoktan seçmeli nihai test maddeleri ortak sonuçlar doğuracağından tek bir puanlayıcı tarafından, hazırlanmış olan yazılı sınav ise üç farklı puanlayıcı tarafından puanlanmıştır. Puanlayıcılara araştırmacı tarafından puanlama cetveli verilmiş ve puanlama için gerekli süre sağlanmıştır. Puanlayıcılar birbirlerini tanımamakta, farklı okullarda görev yapmakta ve farklı kıdem düzeylerinde bulunmaktadır.

### 2.4. Sınırlılıklar

Araştırma 2013-2014 eğitim-öğretim yılı Kars il merkezinde yer alan bir ortaokulda öğrenim gören 8. sınıf öğrencilerinden seçilen 48 kişi ve Fen ve Teknoloji dersi 8. sınıf ' *Hücre Bölünmesi ve Kalıtım'* ünitesi ile sınırlıdır.

## 3. BULGULAR

Bu bölümde araştırmanın alt problemleri için toplanan verilerden elde edilen bulgular, tablo ve açıklamalarıyla birlikte verilerek bunlara dayalı yorumlar yapılmıştır.

*Performansın Ölçülmesinde Kullanılan Çoktan Seçmeli Teste Ait Özellikler:* Çoktan seçmeli test için belirtke tablosuna göre oluşturulan 40 soruluk ön uygulama için betimsel istatistikler Tablo 1' de verilmiştir.

**Tablo 1.** Çoktan Seçmeli Testin Ön Uygulamasına Ait Betimsel İstatistikler

| | |
|---|---|
| Öğrenci Sayısı (N) | 90 |
| Madde Sayısı (K) | 40 |
| Aritmetik Ortalama | 50.16 |
| Varyans ( $s^2$) | 468.85 |
| Standart Sapma (s) | 21.65 |
| En Düşük Puan ( Min.) | 15.00 |
| En Büyük Puan (Max.) | 92.00 |
| Ortalama Güçlük | .523 |
| Ortalama Ayırt Edicilik | .544 |

Çoktan seçmeli testin ön uygulamasından elde edilen madde istatistiklerine göre hazırlanan yazılı sınav sorularının doğrultusunda 22 madde nihai uygulama için seçilmiştir. Çoktan seçmeli test için KR-20 güvenirlik değeri hesaplanmış ve bu katsayının .896 olduğu görülmüştür. KR-20 ile hesaplanan güvenirlik katsayısı testin kendi içinde tutarlılığının bir ölçüsü olup bu değerin yüksek çıkması testin güvenilir olduğu anlamına gelmektedir (Başol, 2016).

Öğrencilere ilk olarak uygulanan çoktan seçmeli test önceden belirlenen kazanımları temsil eden 22 test maddesi ile değerlendirilmiştir. Bunun için öncelikle öğrencilerin doğru cevapları hesaplanmış, 100 üzerinden puanlara dönüştürülmüştür. Çoktan seçmeli teste ait istatistikler Tablo 2' de verilmiştir.

**Tablo 2.** Çoktan Seçmeli Test İle Yapılan Nihai Uygulamaya Ait Betimsel İstatistikler

| Soru Sayısı | n | Ortalama | Medyan | Mod | Mak. | Min. | Ranj | Çarpıklık | Basıklık |
|---|---|---|---|---|---|---|---|---|---|
| 22 | 48 | 68.37 | 68.18 | 68.18 | 100 | 18.18 | 81.82 | -.55 | -.33 |

Uygulanan çoktan seçmeli testte her bir maddeden alınabilecek en düşük puan bir, testten alınabilecek en yüksek puan 22' dir. Puanlar 100 üzerinden değerlendirmeye alınmış ve istatistiksel işlemler bu puanlar üzerinden yapılmıştır. Dönüştürülen puanlara göre çoktan seçmeli testin ortalaması 68.37, medyanı 68.18, modu 68.18' dir. Bu durumda puanların normal dağılım gösterdiğine işaret etmektedir. Testten alınan en yüksek puan 22 sorunun hepsini doğru cevaplayan üç kişi için 100, testten alınan en düşük puan ise dört doğru ile 18.18 olarak hesaplanmıştır. Puanların ortalamasının 50'den yüksek olması öğrencilerin başarı seviyelerini %50 den yüksek olduğunun göstermektedir.

*Alt Problem 1:* 'Çoktan seçmeli test için G Kuramına göre kestirilen parametrelerin varyansları ve toplam varyansları açıklama yüzdeleri nedir?'

Çoktan seçmeli test için birey (b) ve madde (m) değişkenlerinin değişimlerini ve varyans kaynaklarının oranlarını belirlemek için tek değişkenli G (Genellenebilirlik) çalışması yapılmıştır.

**Tablo 3.** Tek Değişkenli G Çalışması Sonucunda Ölçmenin Kestirilen Varyansları ve Toplam Varyansı Açıklama Oranları

| Varyans Kaynağı | Sd | Toplam Kareler | Kareler Ortalaması | Varyans | % |
|---|---|---|---|---|---|
| b | 47 | 45.814 | .975 | .037 | 16.9 |
| m | 21 | 20.235 | .964 | .017 | 7.6 |
| bm | 87 | 162.311 | .165 | .164 | 75.5 |
| Toplam | | | | | 100 |

Tablo 3 incelendiğinde birey (b) ana etkisi için kestirilen varyans bileşeninin (.037) toplam varyansın %16.9' unu açıkladığı görülmektedir. Tek değişkenli modelle yapılan incelemede bireyler için kestirilen varyans bileşeni, toplam varyans içinde en yüksek ikinci paya sahip olan varyans bileşenidir. Genellenebilirlik çalışmalarında, birey ana etkisi evren puanı varyansı olarak değerlendirilir ve ölçülen özellik açısından bireyler arası farklılaşmayı ifade eder (Shavelson ve Webb, 1991; Brennan, 2001). Bireyler için kestirilen varyansın toplam varyans içindeki oranının daha fazla olması istenilen bir durumdur. Bu ölçme ile elde edilen boyutta bireyler arası farklılıkların ortaya çıkarılabildiğinin bir göstergesidir (Güler, 2008).

Madde (m) ana etkisi için tek değişkenli modelle yapılan G çalışmasında kestirilen varyans bileşeni (.017) toplam varyansın %7.6' sını açıklamaktadır. Madde ana etkisinin varyans bileşeni büyüklüğün, toplam varyans değişkeni büyüklüğünde üçüncü ve en az orana sahiptir.

Birey x madde ortak etkisi (.164) toplam varyansın %75.5' ini açıklamaktadır. Birey x madde ortak etkisi tek değişkenli modelle yapılan G çalışmasında elde edilen en büyük varyans

değeridir. Bu durum; bu ölçme için birey x madde ortak etkisinden kaynaklanan farklılığın büyük olduğunu, belli bireylerin bağıl durumlarının bir maddeden diğerine çok farklılaştığını göstermektedir. Ayrıca birey x madde varyans değerinin büyük olması birey ve madde ortak etkisi veya tesadüfü hataların büyük olabileceği anlamına gelebilir.

*Alt Problem 2:* 'Çoktan seçmeli test için yapılan K çalışması sonuçlarına göre G ve Phi katsayılarının değişimleri nasıldır?'

Performansın ölçülmesinde kullanılan çoktan seçmeli test için 22 madde ve madde sayısının arttırılıp azaltılması durumlarında G Kuramı çalışması ile yapılan K çalışması sonucu elde edilen G ve Φ katsayıları Tablo 4' de verilmiştir.

**Tablo 4.** Performansın Ölçülmesine İlişkin Yapılan K çalışması İle Madde Sayıları Senaryolarına Göre G ve Phi Katsayıları

| Madde Sayısı | Φ | G |
|---|---|---|
| 18 | .801 | .785 |
| 20 | .818 | .803 |
| 22 | **.831** | **.817** |
| 24 | .843 | .829 |
| 26 | .854 | .841 |

Tablo 4' te çoktan seçmeli testin madde sayılarının arttırılıp azaltılması durumlarına göre hesaplanan G ve Φ katsayıları verilmiştir. Tabloya göre, madde sayısının nihai testteki değerine göre yapılan analiz sonuçlarına göre; Φ katsayısı .831 ve G katsayısı .817 olarak kestirilmiştir.

Tablo 4 incelendiğinde, madde sayısının azaltılması durumlarında Φ katsayısı ve G katsayılarının azaldığı, madde sayısının arttırıldığı durumlarda Φ katsayısı ve G katsayılarının arttığı gözlemlenmiştir. Ayrıca, 20 madde için elde edilen değerlerin KTK'da Cronbach α değerine karşılık gelmekte ve madde sayısını azaltıp-arttırmanın sonucunda elde edilen güvenirliğin yine KTK'da kestirilebilmekte; G katsayısının avantajı sadece mutlak değerlendirmeler için kullanılabilecek bir güvenirlik değerinin elde edilmesine imkan tanınmasıdır.

*Performansın Ölçülmesinde Kullanılan Yazılı Sınava Ait Özellikler:* Performansın ölçülmesine yönelik uygulanan yazılı sınav 11 maddeden oluşmaktadır. Uygulanan sınav üç farklı puanlayıcı tarafından puanlanmış ve puanlayıcılar üzerinden elde edilen verilerle işlemler gerçekleştirilmiştir. Yazılı sınava yönelik puanlayıcılardan elde edilen puanlara ait betimsel istatistikler Tablo 5' te verilmiştir.

Tablo 5 incelendiğinde, 48 öğrencinin 11 madde üzerinden aldıkları puanlara ilişkin en yüksek ortalama birinci puanlayıcıya aittir ve 56.187 şeklindedir. En düşük ortalama ise 34.708 ile üçüncü puanlayıcıya aittir. İkinci puanlayıcı 45.479 ile puanlayıcı ait ortalama değeri ise bu iki değer arasında yer almaktadır. Birinci puanlayıcıya ilişkin ortanca değer aritmetik ortalamadan yüksektir ve puanların hafif sola çarpık bir dağılım gösterdiği söylenebilir. İkinci ve üçüncü puanlayıcıya ilişkin ortanca değerlerinin aritmetik ortalamadan küçük olması ise puanların hafif sağa çarpık bir dağılım gösterdiğini ortaya koymaktadır. Bu durum çarpıklık katsayılarının birinci puanlayıcıya ait puan değerleri için hafif negatif, ikinci ve üçüncü puanlayıcılara ait puan değerleri

içinse hafif pozitif çıkmasıyla da görülmektedir. Puanlayıcıların verdikleri puan değerlerine ait Cronbach alfa (α) güvenirlik katsayıları birbirine yakın ve yüksek değerlerdir. Puanlayıcıların verrmiş oldukları puan değerlerinin ortalamalarının birbirlerinden farklı olmasının; öğrencilerin sorulara verdikleri yanıtlara açıklık derecelerine göre ya bütüncül ya da ayrıntılı olarak puanlama yapmış olmalarının neden olduğu düşünülmektedir.

**Tablo 5.** Performansın Ölçülmesinde Yapılan Yazılı Sınav İçin Üç Puanlayıcı Ait Betimsel İstatistikler (N=48)

| İstatistikler | 1. Puanlayıcı | 2. Puanlayıcı | 3. Puanlayıcı |
|---|---|---|---|
| Ortalama | 56.187 | 45.479 | 34.708 |
| Medyan | 56.5 | 44.5 | 29.5 |
| Mod | 34 | 43 | 30 |
| Std. Sapma | 2.362 | 2.358 | 2.411 |
| Varyans | 557.943 | 556.170 | 581.360 |
| Çarpıklık | -.184 | .248 | .609 |
| Basıklık | -.970 | -.990 | -.769 |
| Minimum | 8 | 8 | 2 |
| Maksimum | 96 | 83 | 84 |
| α güvenirliği | .850 | .870 | .870 |

*Alt Problem 3:* 'Yazılı sınav için yapılan G Kuramına göre kestirilen parametrelerin varyansları ve toplam varyansları açıklama yüzdeleri nedir?'

Matematik performansının ölçülmesine yönelik hazırlanan 11 maddelik yazılı ölçme aracının G çalışması ile elde edilen varyanslarını ve varyans yüzdelerini hesaplamak için tümüyle çaprazlanmış b x m x p modeli uygulanmıştır. Ölçmenin uygulandığı 48 öğrenci, 11 madde ve üç puanlayıcıdan oluşan verilerde tek değişkenli modelle yapılan G çalışması için; kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri b, m ve p ana etkileri ile bm, bp, mp, ve bmp ortak etkileri Tablo 6' da verilmiştir.

**Tablo 6.** Tek Değişkenli G Çalışması Sonucunda Ölçmenin Kestirilen Varyansları ve Toplam Varyansı Açıklama Oranları

| Varyans Kaynağı | sd | Toplam Kareler | Kareler Ortalaması | Varyans | % |
|---|---|---|---|---|---|
| b | 47 | 5193.657 | 110.503 | 3.082 | 16.0 |
| m | 10 | 44.326 | 4.433 | -.027 | .0 |
| p | 2 | 127.971 | 63.986 | .130 | .7 |
| bm | 470 | 8652.705 | 18.410 | 1.153 | 6.0 |
| bp | 94 | 502.514 | 5.346 | -.873 | .0 |
| mp | 20 | 96.807 | 4.840 | -.211 | .0 |
| bmp | 940 | 14052.708 | 14.950 | 14.950 | 77.4 |
| Toplam | | | | | 100 |

Tablo 6'ya göre, birey (b) ana etkisi için kestirilen varyans bileşenini (3.08173) toplam varyansın %16' sını açıklamaktadır. Tek değişkenli modelle bireyler için kestirilen varyans bileşeni, toplam varyans içinde en yüksek ikinci sırada paya sahiptir.

Madde (m) ana etkisi için kestirilen varyans bileşeni tek değişkenli modelle yapılan G çalışmasına göre kestirilen varyans bileşeni eksi değer aldığı için (-.02686) toplam varyansı açıklama yüzdesi içinde (%0) bir etkiye sahip olmadığı görülmüştür. Varyansın *sıfir* alınmasının nedeni G Kuramı çalışmalarında varyans değerinin negatif çıkması durumlarında uygulanan dört farklı yöntemden biri olmasıdır (Brennan, 2001). Shavelson ve Webb (1981)'e göre negatif varyanslar örnekleme hatalarından ya da yanlış model seçiminden kaynaklanmış olabilir.

Shavelson ve Webb (2005) negatif varyans söz konusu olduğunda dört çözüm önerisi olduğunu belirtmiştir: Cronbach, Gleser, Nanda ve Rajaratnam (1972) negatif varyans değerinin yerine sıfır yazmayı önermişler, ikinci öneri olarak (Brennan, 2001) negatif varyansların sıfır alınmasını ancak beklenen ortalama kareler eşitliğinde negatif varyansların olduğu gibi kullanılmasını, üçüncü öneri ise (Shavelson ve Webb) Bayesian metot kullanılarak tahmin edilen varyans için en küçük değerin sıfır olarak değiştirilmesini, son olarak Searle (1987) maksimum olabilirlik modeli kullanılarak negatif varyansların önüne geçilmesini önermiştir ( Akt. Shavelson ve Webb, 2005). Mevcut çalışmada Cronbach, Gleser, Nanda ve Rajaratnam (1972)'ın önerisi dikkate alınarak negatif varyanslar 0 alınmıştır.

Buna göre, puanlayıcı ana etkisinin G çalışması ile kestirilen varyans bileşeni (.13021) toplam varyansın %0.7' ini açıklayarak toplam varyans içinde dördüncü sırada yer almaktadır. Puanlayıcı etkisinin tek değişkenli modelle yapılan G çalışması ile kestirilen varyans oranının düşük olması, puanlayıcıların tüm bireyler için yaptıkları puanlamalar arasında bir fark bulunmadığını, puanlamalar arasında da bir tutarlılığın olduğunu göstermektedir.

Birey x madde (bm) ortak etkisi (1.15344) toplam varyansın %6' sını açıklamaktadır. Birey x madde ortak etkisi tek değişkenli modelle kestirilen en yüksek üçüncü değere sahip varyans değeridir. Bu da birey x madde ortak etkisinden kaynaklanan farklılığın büyük olduğunu, belli bireylerin bağıl durumlarının bir maddeden diğer maddeye çok farklılaştığını göstermektedir (Güler, 2008).

Birey x puanlayıcı (bp) ortak etkisi (-.87307) toplam varyansın %0' ını açıklamaktadır. Madde x puanlayıcı (mp) ortak etkisi (-.21061) 0' ın altında değer aldığı için toplam varyans içerisinde açıklama yüzdesi %0 dır.

Madde x puanlayıcı etkisinin tek değişkenli modele göre madde x puanlayıcı ortak etkisinden kaynaklanan bir farklılığın olmadığı yorumu yapılabilir.

Birey x madde x puanlayıcı (artık) ortak etkisi varyans bileşenide (14.94969) toplam varyansın %77.4' ünü açıklamaktadır. Bu oran varyans değerleri arasından en büyük değerdir. Birey x madde x puanlayıcı (artık) varyansın büyük olması; birey, madde ve puanlayıcı ortak etkisi veya tesadüfi hataların büyük olabileceğinin bir göstergesi olabilir.

*Alt Problem 4:* 'Yazılı sınav için yapılan K çalışmasına göre farklı senaryolara göre G ve Phi katsayılarının değişimi nasıldır?'

Uygulanan yazılı sınava ait veriler üzerinden madde sayısı ve puanlayıcı sayılarının arttırılıp azaltılması durumlarına göre G Kuramı kullanılarak K çalışması yapılmıştır. Yapılan K çalışmasına ait G ve Φ katsayılarının değişimi Tablo 7' de verilmiştir.

**Tablo 7.** Performansın Ölçülmesine İlişkin Yapılan K çalışması ile Madde ve Puanlayıcı Sayıları Senaryolarına Göre Phi ve G Katsayıları

| Madde Sayıları | Puanlayıcı Sayıları | | | | | |
|---|---|---|---|---|---|---|
| | 2 | | 3 | | 4 | |
| | G | Φ | G | Φ | G | Φ |
| 9 | .763 | .751 | .819 | .810 | .850 | .843 |
| 11 | .797 | .784 | **.847** | **.837** | .874 | .866 |
| 13 | .823 | .809 | .867 | .857 | .891 | .883 |

Tablo 7'ye göre tek değişkenli modelle yapılan ölçme sonuçlarına göre 11 madde ve üç puanlayıcıya göre kestirilen G katsayısı .847 ve Φ katsayısı da .837 olarak kestirilmiştir. Kestirilen katsayı değerlerine bakılarak G katsayısının Φ katsayısından daha yüksek olduğu görülmektedir. Gerek bağıl değerlendirme durumlarında kullanılan G katsayısı ve gerek mutlak değerlendirme durumlarında kullanılan Φ katsayılarının madde sayılarının ve puanlayıcı sayılarının artması durumunda yükseldiği ortaya çıkmıştır. Tüm madde ve puanlayıcı senaryolarında G katsayıları, Φ katsayılarından yüksek değerde çıkmıştır. Madde sayısının aynı kalması durumunda puanlayıcı sayısının artması senaryolarında ortaya çıkan G ve Φ katsayıları; puanlayıcı sayılarının aynı kalması durumunda madde sayısının arttırılması ile kestirilen G ve Φ katsayılarına göre daha yüksek değerlerde ortaya çıkmıştır.

## 4. TARTIŞMA

Araştırma bulgularına göre, bireylerin çoktan seçmeli testten aldıkları puanlar ile yazılı sınavdan aldıkları puanların dağılımlarının paralellik gösterdiği gözlenmiştir. Çoktan seçmeli testten alınan puanların daha yüksek olduğu ortaya çıkan bulgular arasındadır. Ranj değerlerinin değişimine baktığımızda yazılı sınav için her bir puanlayıcının vermiş olduğu puanlar ile çoktan seçmeli teste ait ranj değerinin birbirleri ile çok yakın olduğu görülmektedir.

Yazılı sınav ve çoktan seçmeli test için gerek ortanca gerekse standart sapma değerlerinin ortalama ekseninde değişimleri için belirlenen başarı puanlarının çoktan seçmeli test için dağılımları ile paralellik gösterdiği görülmüştür. Ancak bu araştırmada başarı puanları açısından çoktan seçmeli testten alınan puanların daha yüksek olduğu ortaya çıkmıştır.

Çetin (2009) yapmış olduğu araştırmasında; performans görevi, yazılı sınav ve çoktan seçmeli test arasındaki ilişkiyi farklı değişkenlerle incelemiştir. Çetin, araştırma sonuçlarına göre başarı puanlarının çoktan seçmeli test için daha yüksek olduğu sonucuna ve üç sınav arasında ilişkinin orta düzeyde olduğu sonucuna ulaşmıştır. Ancak ikili ilişkilere bakıldığında çoktan seçmeli test ile yazılı sınav arasındaki ilişkinin daha ileri düzeyde olduğu gözlenmiştir. Diğer ikili karşılaştırmalara göre, yapılmış olan bu araştırmada çoktan seçmeli test ve yazılı sınav arasında ilişki yüksek bulunmuş; uygulama amacına göre sınavların uygulanmasında araştırmacının istediği özelliklere göre her iki sınavında kullanılabilirliği sonucuna varılmıştır. Yazılı sınavda soru sayısının az olması gibi dezavantajlarının yanında puanlayılar arası tutarlılığın sağlanması halinde çoktan seçmeli teste yakın sonuçlar verdiği ortaya çıkmıştır.

Eser (2011) sınav türleri konusunda öğrenci tercihlerini çalışmış olduğu betimsel tarama modelindeki araştırmasında, öğrenciler, başarı puanları daha yüksek olduğu için çoktan seçmeli testleri, yazılı sınavlara göre daha çok tercih ettiklerini belirtmişlerdir. Araştırma sonuçlarına göre

en az tercih edilen sınav türü yazılı sınav türü olarak belirtilmiştir. Yapılan bu çalışmada ise tercih türleri araştırılmamış ancak çoktan seçmeli test puanlarının dağılımlarının yazılı sınav türünden elde edilen puan dağılımlarına göre daha yüksek olduğu sonucuna ulaşılmıştır. Öğrencilerin çoktan seçmeli testlere daha çok aşina olması bu sonuçta etkili olmuş olabilir.

Öğrencilerin bilgilerini kullanarak bir ürün ortaya çıkarılmasını isteyen yazılı sınavların öğrencilerde kaygı ve korkuya neden olduğu ve bu nedenle öğrencilerin başarılarının düşük olduğu farklı araştırmalarda ortaya konulmuştur. Ömür (2002) çalışmasında, öğrencilerin cevap üretmek yerine verilen cevaplar arasından birini seçmeyi daha çok tercih ettiklerini belirtmiştir. Ayrıca başarının yazılı sınavlarda çoktan seçmeli testlere göre daha yüksek olduğu sonucu bu çalışmadaortaya çıkan bir diğer bulgudur.

Bunun aksine bazı çalışmalar da yazılı sınavda ortaya konulan performansın çoktan seçmeli testlere göre daha yüksek olduğu sonucuna ulaşılmıştır. Önder (2008) matematik başarısının ölçülmesi ve sınav kaygı düzeyi üzerine yapmış olduğu çalışmasında; yazılı sınava hazırlanan öğrencilerin başarılarının daha yüksek olduğunu belirtmiştir. Ayrıca çalışmada, hangi tür sorularla sınavlara hazırlanırsa hazırlansınlar, öğrencilerin yazılı sınavlarda daha başarılı oldukları sonucu elde edilmiş; yazılı sorularla sınava hazırlanan öğrencilerin yanı sıra, çoktan seçmeli test sorularla sınava hazırlanan öğrencilerin de yazılı sınavlardan daha iyi bir performans gösterdiği bulunmuştur. Oysa, bu araştırmanın bulgularından birisi öğrencilerin performans puanlarının, çoktan seçmeli sınav için yazılı sınava göre daha yüksek olduğudur. Alan yazın incelendiğinde farklı sınav türlerinin karşılaştırıldığı ve üzerinde G Kuramı çalışması yapılan araştırmalara rastlanmamıştır. Daha çok performansın belirlenmesinde puanlayıcıların birbirleri ile tutarlılığının incelendiği ve farklı desenlere göre karşılaştırılmaların yapıldığı araştırmalar mevcuttur.

Yapılan analizlere göre; G Kuramına göre puanlayıcılara ait puanlayıcı değişkenliğinin etkisinin düşük olduğu ortaya çıkmıştır. Ortaya çıkan bu sonuçların benzer başka çalışmalarda da ortaya çıktığı görülmüştür. Güler (2008) farklı kuramlara göre karşılaştırma yaptığı çalışmasında; matematik başarısını belirlemede uygulanan klasik sınav verileri üzerinden KTK, G Kuramı ve ÇDRM çalışmaları yapmıştır. Elde edilen bulgulara göre G Kuramı çalışması sonuçlarına göre puanlayıcılar arasında tutarlılığı yüksek bulunmuştur. Nalbantoğlu (2009) puanlayıcıların birlikte ve dönüşümlü olarak puanlamalarında sonuçlar arasında paralellik olduğu ve puanlamaların birbirleri ile tutarlı olduğu sonucuna ulaşılmıştır.

LLabre 1978'deki çalışmasında farklı modlar ve farklı yazma becerilerini değerlendiren puanlayıcıların vermiş oldukları puanların aradaki zaman ve farklı ortamlara rağmen tutarlı sonuçlar verdiğini belirtmiştir. Puanlayıcı sayısının artması halinde güvenirlik değerinin yükseldiği sonucu araştırmadan çıkan sonuçlardandır.

Çoktan seçmeli ve yazılı sınavlarda madde sayısının artması sonucu güvenirlik değerinin arttığı bulgularda gözlenmiştir. Puanlayıcı ve madde sayısının artması araştırmanın güvenirliği açısından önemli bir özelliktir. Ancak uygulama, maliyet ve zaman gibi etkenlerden dolayı araştırmalarda hangisinin tercih edilebileceği hakkında bir noktaya varılmak istendiğinde bulgular dahilinde çoktan seçmeli test için madde sayısının arttırılmasının; yazılı sınav için puanlayıcının sayısının arttırılmasının güvenirlik değerlerini daha çok yükselttiği görülmektedir.

Her iki sınav türü içinde güvenirlik çalışması yapılmış ve güvenirlik indeksleri olarak KTK için α ve G Kuramı için G katsayısı hesaplanmıştır. Araştırma için hesaplanan bu değerlere göre α ve G katsayıları oldukça yüksek ve birbirlerine yakın bulunmuştur. Wang (2005) benzer bir çalışmada farklı güvenirlik indekslerini hesaplamış ve karşılaştırmıştır. Çalışmanın sonucunda α ve G katsayısının birbirine yakın olduğu sonucuna ulaşılmıştır.

## 5. KAYNAKLAR


Aiken, L.R. (1991). *Psychological testing and assessment* (7. baskı). Boston: Allyn and Bacon.

Alharby, E.R. (2006). *A comparison between two scoring methods, holistic vs. Analytic using two measurement models, the Generalizability Theory and the many facet Rasch measurement within the context of performance assessment*. Unpublished doctoral dissertation. The Pennsylvania State University Faculty of Education, Pennsylvania.

Alkahtani, S.F. (2012). *Oral performace scoring using generalizability Theory and many-facet Rasch measurement: a comparison study*. Unpublished doctoral dissertation. The Pennsylvania State University, Pennsylvania.

Atılgan, H. (2004). *Genellenebilirlik Kuramı ve çok değişkenlik kaynaklı rasch modelinin karşılaştırılmasına ilişkin bir araştırma*. Yayınlanmamış doktora tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Au, F., Prahardhi, S., & Shiell, A. (2008). Reliability of two instruments for critical assessment of economic evaluations. *Value in Health, 11*, 435- 439.

Başol, G. (2008). Bilimsel araştırma süreci ve yöntem. İçinde Kılıç, O. & Cinoğlu M. (Ed.), *Bilimsel araştırma yöntemleri*, Bölüm 5, İstanbul: Lisans Yayıncılık.

Başol, G. (2016). *Eğitimde ölçme ve değerlendirme*. Genişletilmiş 4. Baskı, Ankara: Pegem Yayıncılık.

Bottema-Beutel, K., Lloyd, B., Carter, E.W. & Asmus, J.M. (2014). Generalizability and decision studies to inform observational and experimental research in classroom settings. *American Journal on Intellectual and Developmental Disabilitie, 119*(6), 589–605.

Brennan, R.L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Brooks, G.P. & Johanson, G.A. (2003). Test Analysis Program. *Applied Psychological Measurement*, 27, 305-306.

Brown, J.D., (2005). Generalizability and decision studies. *SHIKEN: JALT Testing&Evaluation SIG Newsletter. 9*(1), 12 – 16.

Burns, K.J. (1998). Beyond classical reliability: Using Generalizability Theory to assess dependability. *Research in Nursing and Healty, 21*, 83-90.

Burton, E.B. (1998). *An investigation of the school-level generalizability of performance assessment results*. Unpublished doctoral dissertation. Rutgers University, New Jersey.

Çakıcı Eser, D. (2011). *Genellenebilirlik Kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlığın karşılaştırılması*. Yayınlanmamış yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

EduG 2012 software. EduG version 6.1-e, Generalizability Study. Société Suisse pour la Recherche en Education, Groupe de travail Edumétrie – Qualité de l'évaluation en éducation; software prepared by Maurice Dalois and Léo Laroche, Educan Inc., Longueuil, Quebec, Canada.

Gao, X. & Brennan, R.L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14*, 191-203.

Güler, N. (2008). *Klasik Test Kuramı, Genellenebilirlik Kuramı ve Rasch modeli üzerine bir araştırma*. Yayınlanmamış doktora tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Güler, N. (2009). Genellenebilirlik Kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması. *Eğitim ve Bilim, 34*, 154.

Güler, N., Kaya Uyanık, G. ve Taşdelen Teker, G. (2012). *Genellenebilirlik Kuramı*. Ankara: Pegem Yayıncılık.

Kaya, G. (2011). *Genellenebilirlik Kuramının doldurma kavram haritası değerlendirme çalışmasına uygulanması*. Yayınlanmamış yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Moon, S.Y. (1995). *Performance assessment: Measurement ıssues of generalizability, dependability of scoring and relative information on student performance.* Unpublished doctoral dissertation. The Florida State University, Tallahassee.

Mushquash, C. & O'Connor, B. P. (2006). SPSS and SAS programs for Generalizability Theory analysis. *Behavior Research Methods*. *38*(3), 542-547.

Nalbantoğlu, F. (2009). *Performans ölçümlerinde Genellenebilirlik Kuramıyla farklı desenlerin karşılaştırılması*. Yayınlanmamış yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

O' Neill & O' Neill (2015). Improving QST reliability-moreraters, tests, oroccasions? A multivariate generalizability study. *The Journal of Pain, 16*(5), 454-462.

Rentz, J.O. (1987). Generalizability Theory: a comprehensive method for assessing and improving the dependability of marketing measures. *Journal of Marketing Research, 24*(1), 19-28.

Shavelson, R.J. & Webb, N.M. (1981). Generalizability Theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133–166.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory*: *A primer*. Newbury Park, CA: Sage.

Shavelson, R.J. & Webb, N.M. (2005). *Generalizability Theory*. Web: http://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20AERA.pdf adresinden alınmıştır.

Senemoğlu, N. (2002). *Gelişim öğrenme ve öğretim: kuramdan uygulamaya*. Ankara: Gazi Kitapevi.

SPSS Inc. Released 2007. SPSS for Windows, Version 16.0. Chicago, SPSS Inc.

Solano-Flores, G. & Li, M. (2006). The use of Generalizability (G) Theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13-22.

Suen, H.K., & Lei, P.W. (2007). Classical versus Generalizability Theory of measurement. *Educational Measurement, 4*, 1-13.

Tekindal, S. (Ed.) (2011). *Eğitimde ölçme ve değerlendirme*, Ankara: PegemA Yayıncılık.

Turgut, M.F. ve Baykul, Y. (2010). *Eğitimde ölçme değerlendirme*. Ankara: Pegem Yayıncılık.

Wang, Z. (2005). *Estimating reliability under a Generalizability Theory model for writing scores in C-base*. Yayınlanmamış doktora tezi. University of Missouri, Columbia.

Yelboğa, A. (2007). *Klasik Test Kuramı ve Genellenebilirlik Kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi*. Yayınlanmamış doktora tezi. Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

# Summary

## Introduction

Education is the process of transferring past values, moral standards, knowledge and skills to new generations of society as a process of creating the desired behaviors (Senemoğlu, 2002). Through the education, individuals are expected to acquire knowledge, skills and attitudes in certain topics. It is always a matter of curiosity how much the individuals in the training process have benefited from the training or to what extent the learning objectives are achieved. Through the evaluation, it is expected to define the level of knowledge, skills and attainment of attitudes. Therefore, assessment results can be considered as the identificators of the efficiency of the teaching process and expected to enrich the teaching methods that are currently in use in order to get more effectiveness and satisfaction.

Taking into consideration the method used in teaching, among the different measuring instruments, the most suitable one is selected to check whether the learning goals has been achieved or not. The measuring instruments used according to the purpose of the evaluation also vary.

The instrument to be used is decided according to the factors such as the student's readiness, the environment for the examination, time limit and exam conditions. It is worth investigating whether the results obtained with different measuring tools are similar. In this way, it could be possible to know whether the measurement results obtained from different sources are reliable.

In the research studies aimed at measuring the performance of the participants, the researchers used statistical studies using the Classical Test Theory (CTT). CTT is preferred more often because of its ease of use and familiarity. However, with CTT, it is not possible fully understand the inconsistencies in the scores. On the other hand, Generalizability Theory is particularly well suited to distinguish the sources of inconsistencies in observed scores. The Generalizability Theory (G Theory) allows comparison of research results with the reason that it handles the sources of errors at the same time and places them in relation to each other.

The Generalizability (G) Theory is a statistical theory that enables the determination of the reliability of measurement results, the design, the investigation and the conceptualization of reliable observations. Generalizability Theory is an extension of the Classical Test Theory (Cronbach et al., 1972, Brennan, 2001).

According to Shavelson and Webb (1991), Generalizability Theory is an extension of Classical Test Theory from four different perspectives: 1. Generalizability Theory deals with multiple variance sources in a single analysis. 2. It defines each variance source. 3. It allows calculating two different reliability coefficients (G coefficient and phi coefficient) for making relative decisions based on both individual performances as well as absolute decisions about individual performances. 4. Depending on a specific purpose, it is possible to arrange measures (Decision "K" studies) that can reduce the measurement error to the greatest extent possible.

Reliability search in G Theory is conducted in two steps; first is the Generalizability study (G-study) and the second is the Decision study (K-study) (Kaya, 2011). Among these, G study is designed to provide a reasonable and economically versatile isolation and estimation of the measurement error (Shavelson & Webb, 2005). In the G-study process, all sources of variability (variance components) and interactions between them are estimated using the ANOVA method to

generalize the sample to the equation. These predicted variance components are used in the next phase of the K-study (Kaya, 2011).

With the K study, the generalization coefficient (G coefficient) and the reliability index (Phi coefficient), which are similar to the reliability coefficient, are reached. In the G model, the Phi (Φ) coefficient is used with the absolute decision model.

## Methodology

The purpose of the study is to investigate the consistency of the achievement scores obtained from different measurement instruments on the same content. It is also aimed to determine the error amounts of the measurement results obtained in the same individuals, in different situations with different measurement tools separately and for each variable and their combinations with each other. The current study was carried out based on the Theory of Generalizability because different measurement tools and multiple error sources were considered.

The population is the 8th grade students in Kars province during 2013-2014 school years. The study sample, selected through purposeful sampling, composed of 48 students, attending Atatürk Middle School in Kars city center.

The items selected for the multiple-choice test were at different item difficulty levels, the KR 20 reliability value was found to be .90. In order to ensure the validity of the written exam, the questions were prepared and consulted to an area expert and a language expert and the application form was prepared. The items were given to a female and male student beforehand to ensure that there was no misunderstanding or confusion, also to make sure they serve neutrality, both the items of the multiple choice test and the questions in the essay.

## Results and Discussion

According to the measurement results with the univariate model, the G coefficient is estimated as .817 and the Φ coefficient as .831 according to the 22 items included in the multiple choice exam. G coefficient is estimated with respect to three scorers and 11 items are estimated as .847 and Φ coefficient is estimated as .837.

According to the analysis of the variance components for the multiple choice exam; the variance component predicted for the individual has the second highest share, the main effect of the substance has the third and least proportion of the total variance, and the common effect of the individual and the substance has the greatest variance value.

According to analysis of variance components for the written exam; as for the indiviudal component, the variance component was found to be the highest in the total variance and the total variance did not have an effect in the percentage of the explanatory value (0%), as the variance component predicted for the G run with the univariate model predicted for the substance main effect was negative.

The low variance ratio predicted by the G study with the univariate model showed that there was no difference between the scorers for all individuals, therefore there was a consistency between the scorers. Individual x item x scorer (error) source had the largest variance proportion in the common effect variance component.

According to the research findings, it was observed that the distribution of the grades of the individuals in the written exams and the grades they got from the multiple choice test were parallel.

For the written test and multiple choice tests, the median and the standard deviations were found to be consistent with the distributions for the multiple-choice exam.

The results of the analyzes showed that, in terms of CTT, the high correlation between the scorers indicated low scorer effect and according to G Theory the low scorer effect meant high consistency among the scorers. The increase in the number of items of multiple-choice and written exam lead to more reliable scores. The number of items and scorers are important in terms of the reliability of the research. Reliability studies were carried out in both types of tests and Cronbach's alpha (α) for Classical Test Theory and Generalizability for G Theorem were calculated as reliability indices. According to the results, Cronbach's alpha (α) and G coefficients are very high and close to each other.

*Int. J. Asst. Tools in Educ., Vol. 4, Issue 2, (2017) pp. 166-181*

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

*"Research Article"*

# Are We Measuring Teachers' Attitudes towards Computers in Detail?: Adaptation of a Questionnaire into Turkish Culture

**Nilgün Günbaş**[*1] **Özden Demir**[2]

[1]Kafkas University, Faculty of Edu., Dep. of Math. and Sci. Education, Mathematics Education, Kars, Turkey
[2]Kafkas University, Faculty of Edu., Dep. of Educational Sciences, Curriculum and Instruction, Kars, Turkey

***Abstract***

Teachers' perceptions of computers play an important role in integrating computers into education. The related literature includes studies developing or adapting a survey instrument in Turkish culture measuring teachers' attitudes toward computers. These instruments have three to four factors (e.g., computer importance, computer enjoyment, computer confidence) and 18 to 26 items under these factors. The purpose of the present study is to adapt a more detailed and stronger survey questionnaire measuring more dimensions related to teachers' attitudes. The source instrument was developed by Christensen and Kenzek (2009) and called Teachers' Attitudes toward Computers (TAC). It has nine factors with 51 items. Before testing the instrument, the interaction (e-mail) factor was taken out because of the cultural differences. The reliability and validity testing of the translated instrument was completed with 273 teachers' candidates in a Faculty of Education in Turkey. The results showed that the translated instrument (Cronbach's Alpha: .94) included eight factors and consisted of 42 items under these factors, which were consistent with the original instrument. These factors were: *Interest* (α: .83), *Comfort* (α: .90), *Accommodation* (α: .87), *Concern* (α: .79), *Utility* (α: .90), *Perception* (α: .89), *Absorption* (α: .84), and *Significance* (α: .83). Additionally, the confirmatory factor analysis result for the model with eight factors was: *RMSEA*=0.050, $\chi^2/df$=1.69, *RMR*=0.075, *SRMR*=0.057, *GFI*= 0.81, *AGFI*= 0.78, *NFI*= 0.94, *NNFI*=0.97, *CFI*=0.97, *IFI*= 0.97. Accordingly, as a reliable, valid and stronger instrument, the adapted survey instrument can be suggested for the use in Turkish academic studies.

***Article Info***

***Received***
*03 February 2017*

***Revised***
*23 March 2017*

***Accepted***
*28 March 2017*

***Keywords***
Computer Attitude,
Teachers,
Teacher Candidates,

## 1. INTRODUCTION

The use of computers is essential in educational settings. Thus, it is important for teachers to be experienced in computer related skills. Computer literacy courses are one of the required courses in Colleges of Education in the Turkish Universities. Teacher candidates are given computer related skills in these courses as it is necessary to have qualified teachers who know how

[*]Corresponding Author E-mail: ngunbas@gmail.com

2148-7456 /© 2017 DOI: 10.21449/ijate.307973

to deal with computer related problems and keep up with technological developments. In many studies (e.g., Erkan, 2004; Usta & Korkmaz, 2010; Yıldırım & Kaban, 2010; Altun, 2011) it was mentioned that teacher candidates must be equipped with computer technology skills to achieve lifelong learning. Additionally, as a result of their study with teacher candidates and their computer and the Internet use habits, Başol and Çevik (2006) found that teacher candidates must be trained in computer and the Internet use, and necessary adjustments must be provided for them. Additionally they suggested that teacher candidates' current computer and Internet related trainings must be improved. For these reasons, they suggested that it is necessary to provide teacher candidates with technological resources and they must be encouraged to use computers.

Teachers play an important role in integrating computers into education. Hung and Koh (2004) proposed a framework in order to analyze a school's technology integration. In integrating information technologies into schools, there existed four dimensions in socio-cultural factors of schools: school set-up, classroom dynamics, students' behaviors and teachers' attitudes (Hung & Koh, 2004). The authors argue that teacher attitudes affect classroom and student behaviors, and reaching educational goals.

Attitude could be defined as a person's mental and neural readiness affecting their responses to a situation (Khine, 2001 in Erkan, 2004). It can be attributed to a person and that person's tendency to form his/her feelings, thoughts and behaviors about another person or an object (Kağıtçıbaşı, 2016). Attitudes can be shaped and learned with experience (Ekici, Uzun & Sağlam, 2010), directs our behaviors and are the psychological characteristics behind our behaviors (Tavşancıl 2014). Thus it is important to measure it in terms of individuals and community. A person's attitude towards computes, therefore, affects his computer use. Thus, it is highly possible that teachers' positive attitude towards computers is important in organizing educational settings (Aypay & Özbaşı, 2008; Cüre & Özdener, 2008). As time go by so do technological developments. Thus, teachers' perceptions about technology are reported getting more positive parallel to these developments (Cüre & Özdener, 2008). Additionally, Slough and Chamblee (2000) claim teachers, who have witnessed the positive effect of technology in their teaching activities, won't avoid taking advantage of technology.

The more one have experience in using computers, the more he or she has positive attitudes towards computers (Kinzie & Delcourt, 1991; McInerney, McInerney & Sinclair, 1994; Levine & Donitsa-Schmidt, 1998, Deniz 2000; Erkan, 2004; Cüre & Özdener, 2008; Ekici, Uzun & Saglam, 2010; Lehimler, 2016). Those who don't have enough experience in computers might develop negative attitudes towards them (Hashim & Mustapha, 2004). Mitzner et al. (2016) argued that one's attitude towards and positive experience in technology is highly related to her view of technology in terms of its usefulness and ease of use. Teo (2009) argues that teacher candidates' perceptions related to computers is explained by perceived usefulness and perceived ease of use. Cognitive attitude, awareness, and application software ability are some of the predictors for teachers' computer use (Kay, 1990). In a recent study by Teo, Milutinović & Zhou, (2016) found that attitudes towards computers are highly related to perceived usefulness, perceived ease of use, and technological complexity. How proficient one sees himself in using computers is highly related to his attitudes towards computers (Deniz & Köse, 2003). Having a computer home (Çelik & Bindak, 2005; Mumcu & Usta, 2014), and perceptions about the proficiency in computer use (Deniz, 2000) are seen positively effective in teachers' attitudes towards computers. Teacher candidates' attitudes towards computer-based education and computers are found to be positively and significantly related (Oğuz, Ellez, Akamca, Kesercioğlu & Girgin, 2011).

Aypay and Özbaşı (2008) investigated teachers' perceptions about the computer use in schools. As a result of their studies, teachers claimed that the number of computers is not enough in schools, more in-service training about computers must be provided, and teachers must be encouraged for the use of computers in their classes. In their study, Bahar and Kaya (2013) found the following comparisons regarding computer use: Female students are more anxious than male students; those who don't own computers are more anxious than those who own computers; those who easily reach computers are less anxious than those who don't. Moreover, those people with more anxiety about computers see themselves inadequate in solving technology related problems.

Çavuş and Gökdaş (2006) found that the use of computers among teacher-candidates is insufficient, there is no relationship between their gender and the frequency they use computers, and the reason they use the Internet is mostly to find information. Gender and computer ownership are not seen as an effective issues for Turkish teacher-candidates attitudes towards computers (Şahin & Akçay, 2011). However, the year of school a teacher-candidate is in is reported effective on being more/less positive about computer related education.

Determining teachers' beliefs and their attitudes towards computers is important. It was argued that having positive attitudes and beliefs about computers are necessary to be developed in a positive way (Güzeller, 2011). Rana (2012) argues that teachers must have positive attitudes towards computers because their intention for computer use is highly related to their thoughts of their success in integrating technology into their classrooms. Teachers' attitude towards computers is a strong predictor of their attitudes towards using the Internet, as well (Bahar, Uludağ & Kaplan, 2009; Ozden, Aktay, Yilmaz, Ozdemir, 2007). Mumcu and Usta (2014), in their studies, found that teacher candidates use the Internet for research and homework purposes. Teacher candidates, who have positive attitudes towards the Internet, are reported using the Internet often and every day.

There are some computer attitude survey instruments adapted from other cultures into Turkish culture (e.g., Berberoğlu & Çalıkoğlu, 1991; Demir & Yurdugül, 2014) as well as the ones developed in Turkish (e.g., Aşkar & Umay, 2001; Bindak & Çelik, 2006; Yeşilyurt & Gül, 2007). For example, Berberoğlu and Çalıkoğlu (1991) in their studies adapted a survey instrument, which includes three factors, developed by Loyd and Gressard (1984) in the USA. This survey instrument originally included 40 items which were grouped under the following factors: *computer liking* (10 items), *computer confidence* (10 items), *computer anxiety* (10 items) and *computer usability* (10 items). For the validity and reliability of the instrument, they tested the instrument with 282 students. While the factor loads ranged from .77 to .85, the Cronbach's values for the whole scale was .90, for the computer anxiety it was .57, for the computer confidence it was .72, for the computer liking it was .68 and finally for the computer usability it was .72. They found that the adapted survey included only one factor based on Turkish culture and all the factors in the original survey were not observed in the adapted version. As a result, this survey is not strongly sufficient for testing teachers' attitudes towards computers in Turkey. Demir and Yurdugül (2014) adapted a survey instrument which was originally developed by Knezek, Christensen and Miyashita (1998). This instrument included eight factors with 65 items. However, Teo (2008) used only three factors with 20 items from this original instrument and tested it with 183 students in Singapore. Demir and Yurdugül (2014) used the one which Teo (2008) has used. The factors in this instrument were *computer importance* (6 items), *computer enjoyment* (6 items) and *computer anxiety* (8 items). With the Likert scale answers from strongly disagree to strongly agree, they tested the

validity and reliability of the instrument with 1678 students. As a result, they found that the adapted survey including three factors were reliable and valid for Turkish culture.

As for the ones, which were created in Turkish, Yeşilyurt and Gül (2007) developed a computer attitude scale including three factors with 26 items. The factors included *available resources, computer-use ability* and *level of computer use in schools.* Their Cronbach's Alpha for the whole scale was .90. Additionally, Bindak and Çelik (2006) developed a scale measuring primary school teachers' attitudes towards computers. The scale included four factors with 22 items. These four factors were reported as explaining 53.8% of the total variance. Cronbach's Alpha for this scale was .91.

In this study, to present an alternative and a stronger measurement instrument to measure teachers' attitude towards computers, we used a questionnaire instrument with nine subscales with high reliability values ranged from .84 to .94. It is called *the Teachers' Attitudes toward Computers (TAC) Questionnaire Instrument,* created and developed by Christensen and Knezek (2009). The reason to select this questionnaire was to use a stronger scale to measure Turkish teachers' attitudes towards computers. Because it had more factors and more items than other questionnaires in Turkish literature (e.g., Aşkar & Umay, 2001; Bindak & Çelik, 2006; Yeşilyurt & Gül, 2007; Demir & Yurdugül, 2014), we believed that it would bring up more details about teachers' beliefs towards computers. Additionally, it contained much more detailed dimensions in computer attitudes, which is different from other questionnaires.

## 2. METHOD

### 2.1. Sample and Study Design

This study used a quantitative design method. The translation of the survey items into Turkish, item equivalency evaluation, and construct validity testing with exploratory and confirmatory factor analysis were completed in the adaptation process. The study was conducted with 273 teacher candidates from three departments in a Faculty of Education in Turkey. The departments were Elementary School Mathematics Teaching, Turkish Teaching and, Guidance and Psychological Counseling departments. The sampling method for selecting the participants was probability sampling. In this sampling method, the subjects have an equal chance of being selected (McMillan, 2012). A small percent of the population would yield a precise description of the population according to this method. After randomly selecting the participants from three departments, the study was processed.

### 2.2. The Survey Instrument

The Teachers' Attitudes toward Computers (TAC) Questionnaire was created and developed by Christensen and Knezek (2009). In developing the instrument, Christensen and Knezek (2009) have recruited 284 items under 32 subscales from 14 well-valid survey instruments. First of all, an exploratory factor analysis was administered to 621 educators on this version of the instrument. The results showed that 7-factor, 10-factor and 16-factor possible factor structures could be representing teachers' attitudes towards computers. A content analysis revealed that the 7-factor structure was the one that was appropriate. These factors, with the Cronbach's Alphas ranged from .85 and .98, were: *Enthusiasm/ enjoyment, anxiety, avoidance/acceptance, email for classroom learning, negative impact on society, productivity and semantic perception of computers.* They also conducted parallel forms reliability test on these factors by creating A and B forms of the instrument. The reliability results ranged from .85 to .96 in the form A and from .85 to .95 in the

form B. As a result they had 90 items from the results of the parallel forms reliability test in addition to 16 other items measuring teachers' attitude towards computers. These 106 items were then tested with an exploratory factor analysis in two refinement phases: The first phase was held between the years of 1995 and 1997 ($n$ = 621) and the second phase was held between the years of 1997 and 1998 ($n$=1296). As a result, they created a scale with 85 items. The Cronbach's Alpha values for the first phase were as followings: For Interest (9 items) it was .88, for Comfort (8 items) it was .94, for Accommodation (11 items) it was .86, for Interaction (e-mail) (10 items) it was .95, for Concern (10 items) it was .84, for Utility (10 items) it was .89, for Perception (7 items) it was .92, for Absorption (10 items) it was .89, for Significance (10 items) it was .84. In the second refinement phase, they reached to a structure with 85 items. In this structure, the Cronbach's Alpha values for the second phase were as followings: For Interest (9 items) it was .90, for Comfort (8 items) it was .92, for Accommodation (11 items) it was .86, for Interaction (e-mail) (10 items) it was .95, for Concern (10 items) it was .86, for Utility (10 items) it was .92, for Perception (7 items) it was .93, for Absorption (10 items) it was .88, for Significance (10 items) it was .86. As a result of the latest factor analysis conducted in 2000, the final version (i.e., version 6) of the TAC instrument ended up having 51 items.

In 2000, the final version of the instrument (i.e. version 6) was applied to 546 teachers and had reliability values ranged from .84 to .96. These Cronbach's values were as followings: For Interest (5 items) it was .90, for Comfort (5 items) it was .94, for Accommodation (5 items) it was .88, for Interaction (e-mail) (5 items) it was .94, for Concern (8 items) it was .89, for Utility (8 items) it was .90, for Perception (5 items) it was .96, for Absorption (5 items) it was .89, for Significance (5 items) it was .84. In 2003, additionally, this instrument was retested with 786 pre-service teachers and the reliability results ranged from .84 to .94. With 306 in-service teachers, the reliability results ranged from .86 to .97. In 2006, this instrument was retested with K-12 teachers and the reliability results ranged from .89 to .95. In 2008, the reliability test, with 273 pre-service teachers in Texas and Maine, resulted in the range from .87 to .95. This instrument was adapted into other languages as well. For example, it was applied in Mexico in 2006 by Morales and the reliability results ranged from .74 to .98.

The confirmatory factor analysis administered in 2003 on the TAC with 51-item to 1176 teachers from elementary school (%49), middle school (%22), and high school (%29) in Texas, the USA. Goodness-of-fit values were as supported by the goodness of fit index (Tabachnick & Fidell, 2001) *RMSEA* = .048, *SRMR* = .0452, *CFI* = .984.

The original instrument as mentioned earlier has 51 items under the factors of *Interest, Comfort, Accommodation, Interaction (e-mail), Concern, Utility, Perception, Absorption,* and *Significance*. It was necessary to decide whether *the Interaction (e-mail)* factor in the questionnaire has a place in Turkish culture. For this reason, the e-mail factor was judged by a semi-structured interview form with 5-items developed by the researchers. This form was administered to an academician and a teacher, whose area of expertise is Computer Education and Instructional Technology. A content analysis was used in identifying the interview questions. In determining the intercoder reliability, Reliability = number of agreements/ (total number of agreements + disagreements) formula (Miles and Huberman, 1994) was used, and it was found to be .80. In the content analysis, themes and codes were composed. As a result, it was found that e-mail is not used effectively in Turkish culture. The themes and the codes revealed from the interviews with the academician, (i.e., K1) and the teacher (i.e., K2) were as followings:

In the first theme "The effectiveness of e-mail use in education process" and for *the subject differences* code in this theme, K1 reveals that "As I mentioned earlier, students prefer communicating and sharing contents on social media rather than e-mail". K2 states that "e-mail is in no way in use between teachers and students, school management and teachers, and among teachers"

In the second theme "Providing better educational experiences with e-mail use" and for "the official purposes use" code in this theme, K1 claims that "because I think that e-mail is mostly used for official purposes". K2 tells that "e-mail is for data sharing. How could it be used for classes?"

In the third theme "Making education process more interesting with e-mail use" and for "the Internet access problem" code in this theme, K1 states "students, who do not have or have limited internet access, have difficulty with sending e-mails". While for the "students' incapability" code K2 claims "students don't know what e-mail is, what it is used for although they use it to log in to Facebook, Instagram and Twitter. They don't know it could be used for sharing files"

In the fourth theme "Providing more learning opportunities in education process" and for "the internet connection difficulty" code, K1 claims that "if only internet access problem is solved, it might help". For the "lack of interactive content and teacher incapability" code K2 states that "It wouldn't have interactive content. Nothing has come to my mind. It might be my incompleteness".

Lastly in the fifth theme "Increasing motivation with e-mail use in education process" and for "the use of social media" code K1mentions that "Moreover there is Edmodo that I use for educational purposes. It is a social media platform and much more like Facebook. I add my students into the groups in this platform". For "the lack of alternative apps" code, K2, by talking about the EBA system, developed by the Ministry of National Education in Turkey, mentions that "for teachers to communicate with students there is no longer need for dealing with e-mail. The EBA system does and covers everything.

For this reason, the e-mail factor was removed from the questionnaire since it is not in use by educators for education purposes. For the future studies, it is necessary to include more up-to-date social platforms (e.g., cloud storages) in the questionnaire. As a result, because the use of e-mail is not as frequently used in Turkey, *the Interaction (e-mail)* factor was eliminated from the TAC and a 42-item version was used in the present study.

### 2.3. Data Analysis

In the scope of validity testing, exploratory factor analysis (EFA) was used to investigate the construct validity to evaluate the structure of the adapted survey in Turkish culture. In addition, an item-total correlation was calculated to evaluate the strength of the survey in differentiating those with high and low levels. An item analysis was conducted based on the average level of upper and lower groups. Additionally, a Cronbach's Alpha correlation coefficient was calculated to test the consistency of the survey items. A test-retest reliability analysis was also used to test the stability of the survey.

## 3. FINDINGS

Studies on survey instrument adaptation aim adapting a survey, developed in a culture, into different languages and cultures. There are many national and international studies focusing on adaptation surveys in the literature. These studies give information about the survey adaptation

process. In this study, the following phases, suggested by and Hambleton & Bollwark (1991), Hambleton & Kanjee (1993) and Savaşır (1994) were completed: The translation of the items, item equivalency evaluation, and reliability and validity testing of the Turkish translated form.

### 3.1. Translation of the Survey Instrument

As Savaşır (1994) states for the translation of the survey instrument, which is the most important part in adapted survey studies, translators should know both languages and the subject area well, and have experiences in both cultures For this reason, the translation of the instrument, from English to Turkish (i.e., from the source language to target language), was completed by an assistant professor who meets these criteria.

### 3.2. Item Equivalency Evaluation

Upon the completion of the translation, judgmental and statistical techniques were used in order to judge the source and translated instruments in terms of equivalency. In this study, single-translation method was used as a judgmental method. The most important reason to use this method was to investigate and evaluate the item equivalency in the target language. Thus, appropriate expressions in the target language might be chosen and adapted, so that intended meaning of the source language might convey the accurate meaning (Hambleton & Bollwark, 1991).

As one of the judgmental method, back-translation method investigates item equivalency in the source language. In this method, the translated instrument is translated back into the source language and compared to the source instrument. However, because the comparisons are made in the source language, the problems in the target language may not be determined enough (Savaşır, 1994). Additionally, in the back-translation method comprehensibility of the instrument is not taken into account. However, in the single-translation method how participants interpret the instrument can be determined. Therefore, because the back-translation method falls short (Hambleton & Kanjee, 1993; Savaşır, 1994), the single-translation method was preferred in this study.

The first version of the translated form was evaluated in terms of words, terms and expressions, and then compared to the source language. Then, necessary corrections were made to make it appropriate for the target culture. In addition, the Turkish translated draft form was evaluated in terms of Turkish linguistic by a Turkish philologist. Based on the experts' views, the survey items were evaluated one by one and all the necessary alterations were made.

Then, four graduate students from the Curriculum and Teaching department were asked to read and evaluate the form in terms of clarity and suitability. The researchers asked them what each item means to get data on item equivalency. Based on their comments, necessary corrections were made on the items. Additionally, linguistic equivalence was evaluated in terms of consistency between the source and the translated survey instruments (Hambleton & Bollwark, 1991). For this, 40 students from a Department of English were administered with the instruments. They took the English version and then the Turkish version of the instrument over two-week period, respectively. As a result, there was a strong positive relationship between the instruments ($r = 0.90$, $p < .05$).

### 3.3. Validity Testing: Construct Validity

Exploratory factor analysis was performed to examine the Turkish translation of the survey instrument in the frame of Turkish culture. In the exploratory factor analysis, the purpose is to bring variables together to find out new significant factors based on the relationships between the variables (Büyüköztürk, 2002). That is, in order to measure an unknown structure the results of

the scale are taken into consideration to explain the related structure. According to Deniz (2007), exploratory factor analysis is a technique to reveal the dimensions of an adapted scale in the new culture. Thus, this study was completed to determine the TAC's categories, under which the items in the Turkish form fit in. Additionally, the factor loadings of the items were investigated with regard to the scale structure in Turkish culture. Moreover, the Principle Component Analysis, which is often used in social sciences, is used as a factoring technique in the exploratory factor analysis. To reset the correlation between the factors and thus to enable the interpretation of the factors, a Varimax orthogonal rotation was performed. The lower limit was set to 1.00 for the item eigen values to determine the number of factors (Tabachnick & Fidell, 2001; Büyüköztürk, 2002).

The sample size was taken into consideration for the exploratory factor analysis. The sample size was 273 for this study. Before testing the factor analysis, the data was examined in terms of appropriateness for a factor analysis. For this, a Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Barlett's test of Sphericty were performed. The KMO was used to assess the adequacy of the sample size. A KMO value might be between 0 and 1 with the following labels: 0.90 to 1.00 is marvelous, 0.80 to 0.89 is meritorious, 0.70 to 0.79 is middling, 0.60 to 0.69 is mediocre, 0.50 to 0.59 is miserable and 0.00 to 0.49 is unacceptable ( Tabachnick & Fidell, 2001; Cohen, Cohen, West & Aiken, 2003). In addition, if Bartlett's Test of Sphericity value is significant, then the sample size is considered as adequate for the factor analysis. Also, this test shows whether the correlation matrix is appropriate (Tabachnick & Fidell, 2001; Büyüköztürk, 2002). The results suggested that both values are appropriate for a factor analysis. (KMO =.903; Barlett's Test of Sphericty $\chi 2=6.820$ df =861 $p<.001$).

The scale included 42 items under 8 factors. As a result of applying the scale to 273 students, Cronbach's Alpha for the scale in total was found to be .94. For the sub-factors the Cronbach's alpha values were: .90 for the first sub-factor (Utility) (7 items), .90 for the second sub-factor (Comfort) (5 items) , .89 for the third sub-factor (Perception) (5 items) , .84 for the fourth sub-factor (Absorption) (5 items) , .87 for the fifth sub-factor (Accommodation) (5 items), .79 for the sixth sub-factor (Concern) (6 items), .83 for the seventh sub-factor (Significance) (4 items), and .83 for the eighth sub-factor (Interest) (5 items). Preliminary results for the factor analysis indicated that there were ten components with eigen value above 1.00. The scree plot for the eigen values showed that the most important break points were in the eighth factor. In deciding the total number of factors, the eigen value, the percentage of contribution and the scree plot were three criteria that were used the most (DeVellis, 2003). It was argued that the number of factors to the point, where the scree plot takes a horizontal shape, could be used as criteria to specify the appropriate number of factors (DeVellis, 2003).

In addition, the original scale has nine sub-factors. However, the e-mail sub-factor was taken out because of the cultural differences. Thus, the factor analysis for the scale with eight sub-factors (i.e., F1: Utility, F2: Comfort, F3: Perception, F4: Absorption, F5: Accommodation, F6: Concern, F7: Significance, and F8: Interest) were re-applied.

Table-1 shows the structure with eight factors, which was obtained after the factor analysis with two iterations. The factors, which were obtained from the reliability analysis, factor loadings, factor eigen values, percentage of variance, which was explained by the factors, and the Cronbach's Alpha values were included in the table. Additionally, it shows the revised item-total correlations (*r*), common variances and t-values.

**Table 1.** Factors, Factor Loadings, Percentage of Variances Explained by Factors, and Item-Total Correlations Values (r) r: item-total correlations. * Significant at .05 level

| Item # | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | $\bar{X}$ | T | SS | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| m26 | .80 | | | | | | | | 4.44 | -7.62 | .87 | .48* |
| m25 | .78 | | | | | | | | 4.23 | -8.96 | .75 | .53* |
| m24 | .77 | | | | | | | | 4.03 | -8.90 | .82 | .56* |
| m31 | .71 | | | | | | | | 4.20 | -7.84 | .81 | .52* |
| m28 | .71 | | | | | | | | 3.83 | -8.43 | .89 | .49* |
| m27 | .70 | | | | | | | | 4.07 | 8.53 | .88 | .49* |
| m30 | .66 | | | | | | | | 4.13 | -7.60 | .80 | .52* |
| m6 | | .79 | | | | | | | 4.05 | -6.91 | .99 | .47* |
| m9 | | .79 | | | | | | | 4.12 | -7.61 | 1.02 | .53* |
| m7 | | .76 | | | | | | | 4.09 | -9.52 | .96 | .59* |
| m8 | | .75 | | | | | | | 4.21 | -8.62 | .94 | .59* |
| m10 | | .62 | | | | | | | 4.30 | -7.82 | .88 | .59* |
| m34 | | | .87 | | | | | | 4.78 | -9.34 | 1.79 | .42* |
| m35 | | | .86 | | | | | | 4.90 | -10.21 | 1.88 | .49* |
| m33 | | | .85 | | | | | | 4.62 | -8.79 | 1.84 | .44* |
| m36 | | | .80 | | | | | | 4.48 | -9.14 | 1.84 | .42* |
| m32 | | | .63 | | | | | | 5.37 | -11.81 | 1.81 | .56* |
| m40 | | | | .72 | | | | | 3.05 | -9.56 | 1.12 | .53* |
| m38 | | | | .72 | | | | | 3.42 | -6.76 | 1.10 | .40* |
| m37 | | | | .72 | | | | | 3.11 | -11.24 | 1.08 | .55* |
| m42 | | | | .72 | | | | | 3.16 | -7.34 | 1.07 | .42* |
| m39 | | | | .71 | | | | | 3.32 | -8.24 | 1.02 | .50* |
| m13 | | | | | .73 | | | | 4.60 | -6.23 | .71 | .54* |
| m11 | | | | | .72 | | | | 4.52 | -7.68 | .84 | .57* |
| m12 | | | | | .71 | | | | 4.41 | -7.99 | .86 | .50* |
| m14 | | | | | .64 | | | | 4.65 | -5.53 | .66 | .51* |
| m15 | | | | | .57 | | | | 4.32 | -8.44 | .90 | .62* |
| m20 | | | | | | .73 | | | 2.73 | -6.94 | 1.13 | .34* |
| m21 | | | | | | .71 | | | 3.36 | -6.90 | 1.08 | .39* |
| m23 | | | | | | .71 | | | 3.30 | -6.21 | 1.12 | .36* |
| m18 | | | | | | .68 | | | 3.38 | -7.01 | 1.17 | .39* |
| m19 | | | | | | .66 | | | 2.73 | -5.08 | 1.07 | .28* |
| m17 | | | | | | .56 | | | 3.07 | -5.57 | 1.10 | .31* |
| m45 | | | | | | | .73 | | 4.25 | -7.34 | .83 | .50* |
| m46 | | | | | | | .73 | | 4.19 | -6.84 | .90 | .45* |
| m44 | | | | | | | .72 | | 4.39 | -7.73 | .83 | .52* |
| m43 | | | | | | | .52 | | 4.09 | -7.58 | .84 | .53* |
| m4 | | | | | | | | .65 | 3.95 | -9.57 | 1.04 | .61* |
| m2 | | | | | | | | .65 | 4.18 | -7.41 | .89 | .52* |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| m1 | | | | | | | | .60 | 4.00 | -10.14 | 1.03 | .64* |
| m3 | | | | | | | | .59 | 3.10 | -8.30 | 1.14 | .45* |
| m5 | | | | | | | | .54 | 4.18 | -8.07 | .83 | .55* |
| Rank | .66-.80 | .62-.79 | .63-.87 | .71-.72 | .57-.73 | .56-.73 | .52-.73 | .54-.65 | 2.73-5.37 | -11.81--5.08 | .66-1.88 | .28-.64 |
| | | | | | | | | | | | | **Total** |
| *Variance %* | 11.41 | 9.25 | 8.56 | 8.46 | 8.23 | 7.66 | 6.18 | 5.95 | | | | 65.38 |
| *Cronbach's Alpha* | .90 | .90 | .89 | .84 | .87 | .79 | .83 | .83 | | | | .94 |

*Note: To make it easier to follow, factor loadings lower than .30 are not given in the table. F1: Utility, F2: Comfort, F3: Perception, F4: Absorption, F5: Accommodation, F6: Concern, F7: Significance, and F8: Interest*

## 4. DISCUSSION

The factor structure of the TAC was investigated with exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The purpose of exploratory factor analysis is to explore factor structure regarding to the relationship between the variances. Confirmatory factor analysis, investigating the model-data compatibility, tests the hypothesis in regard to the variables (Tabachnick & Fidell, 2001).

The first factor, F1, labeled as "Utility", includes 7 items (i.e., i26, i25, i24, i31, i28, i27 and i30). For example, one item in this factor is "Computer can help me learn". Factor loading within the F1 factor is between .66-.80 and item-total correlation is between .48-.56. Cronbach's Alpha value is .90 for this factor.

The second factor, F2, labeled as "Comfort", includes 5 items (i.e., i6, i7, i8, i9, i10). One example item in this factor is "Working with a computer makes me feel tense and uncomfortable". Factor loading within the F2 factor is between .62-.79 and item-total correlation is between .47-.59. Cronbach's Alpha value is .90 for this factor.

The third factor, F3, labeled as "Perception", includes 5 items (i.e., i32, i33, i34, i35, i36). In this factor the items included adjective-pairs that could explain one's feelings for computer use (e.g., unplesant-plesant). Factor loading within the F3 factor is between .63-.87 and item-total correlation is between .42-.56. Cronbach's Alpha value is .89 for this factor.

The fourth factor, F4, labeled as "Absorption", includes 5 items (i.e., i37, i38, i39, i40, i42). One example item in this factor is "I like to talk to others about computers". Factor loading within the F4 factor is between .71-.72 and item-total correlation is between .40-.55. Cronbach's Alpha value is .84 for this factor.

The fifth factor, F5, labeled as "Accommodation", includes 5 items (i.e., i11, i12, i13, i14, i15). As an example, one item in this factor is "Studying about computers is a waste of time". Factor loading within the F5 factor is between .57-.73 and item-total correlation is between .50-.62. Cronbach's Alpha value is .87 for this factor.

The sixth factor, F6, labeled as "Concern", includes 6 items (i.e., i17, i18, i19, i20, i21, and i23). "Computers dehumanize society by treating everyone as a number" is one of the items in this factor. Factor loading within the F6 factor is between .56-.73 and item-total correlation is between .28-.39. Cronbach's Alpha value is .79 for this factor.

The seventh factor, F7, labeled as "Significance", includes 4 items (i.e., i43, i44, i45, i46). One example item in this factor is "Students should understand the role computers play in society". Factor loading within the F7 factor is between .52-.73 and item-total correlation is between .45-.53. Cronbach's Alpha value is .83 for this factor.

The eighth factor, F8, labeled as "Interest", includes 5 items (i.e., i1, i2, i3, i4, i5). As an example, one item in this factor is "I want to learn a lot about computers". Factor loading within the F8 factor is between .54-.65 and item-total correlation is between .45-.64. Cronbach's Alpha value is .83 for this factor. As a result of the analysis, 5 items were eliminated from 47 items in the translated Turkish scale. The items related to the email factor were removed from the questionnaire with 51 items for the reasons stated above. For this reason, we started to the analysis with 47 items. 29$^{th}$ and 47$^{th}$ items were removed from the analysis after the first phase of the exploratory factor analysis since they did not fit under the *Utility* and the *Significance Factors*, respectively. Similarly, 16$^{th}$ and 22$^{nd}$ items were removed from the *Concern Factor*. The 41$^{st}$ itemwas also removed from the analysis because its factor loading was under .30. Accordingly, the draft scale ended up with having 42 items.

65.38 % of the variances were explained by eight sub-factors. The Cronbach's Alpha for the TAC scale in total was .94. The stability and consistency between the two halves were calculated with Guttman and Split Half test. As a result, the values were .83 for the first sub-factor, .84 for the second sub-factor, .83 for the third sub-factor, .82 for the fourth sub-factor, .79 for the fifth sub-factor, .77 for the sixth sub-factor, .85 for the seventh sub-factor and .80 for the eighth factor. For the whole scale it was .75.

As it can be seen in the Table 1, factor loadings for the entire survey was between .52-.87. For the items, which fit in a certain sub-factor, the factor loadings are generally greater than and equal to .30 in fitting in related sub-factors.

The arithmetic means and the standard deviations for the 42 items ranged from 2.73 to 5.37, and .66 to 1.88, respectively. The participants' total scores were sorted in ascending order to form the top 27% and the bottom 27%. These two groups were labeled as upper and lower groups. These groups were then compared to each other to make sure that the items of the survey differentiate these two from each other. As a result, all the items were found to be significantly differentiating these groups ($p$<.001).

The confirmatory factor analysis was used to test the correctness of the survey with eighth sub-factors. The most common statistical tests to evaluate model fit are $\chi^2$, $\chi^2$/df, RMSEA, NNFI, CFI and GFI (Sümer, 2000; Hoe, 2008; Çokluk, Şekercioğlu & Büyüköztürk 2012). A chi-square test of model-data fit was performed to determine whether the model with eight factors was appropriate. The results were found to be statistically significant for the model-data fit ($\chi^2$=1338.53, sd= 791, p<.01). As a result of the confirmatory factor analysis, the goodness of fit index for the model with seven factors was: *RMSEA*=0.050, $\chi^2/df$=1.69, *RMR*=0.075, *SRMR*=0.057, *GFI*= 0.81, *AGFI*= 0.78, *NFI*= 0.94, *NNFI*=0.97, *CFI*=0.97, *IFI*= 0.97. Thus, these results were compatible with the suggested criteria. The standardized coefficients indicating the relationship between the items and the factors ranged from .28 to .64 and all the items were found to be statistically significant ($p$<.01).

In general, the model showed a perfect fit to the data (*RMSEA*=0.050, $\chi^2/df$=1.69) as supported by the goodness of fit index (Tabachnick & Fidell, 2001; Dorman & Knightley, 2006).

### 4.1. Test-Retest Reliability

Test-retest reliability is a measure showing the stability of a test overtime (Çokluk et al., 2012). Thus in this study, the consistency of Turkish version of the survey is measured with this method. To determine the test-retest reliability coefficient, 60 students from the Faculty of Education were administered with the survey twice over a two-week period. Pearson's Correlation coefficient results showed that there is a strong positive relationship between the test results ($r$=.85, $p$<0.5). It can be concluded that the adapted test is stable and reliable.

## 5. RESULTS

Knowing teacher candidates' attitudes towards computers may contribute to their educational process. The original instrument, Teachers' Attitudes toward Computers (TAC), has nine factors. By taking cultural differences into account, the email factor was eliminated in this study. As a result, the instrument with eight factors was adapted into Turkish culture. As a result of the exploratory factor analysis, Kaiser-Meyer-Olkin (KMO) coefficient and Barlett Sphericty test results were found to be statistically significant.

The confirmatory factor analysis, performed for investigating the compatibility of the model with the collected data and a Chi-Square value, calculated for investigating model-data compatibility were found to be statistically significant. The results of the confirmatory factor analysis for the model with eight sub-factors were appropriate with the suggested criteria. Standardized coefficients, indicating the relationships between the items and relevant factors, ranged from .28 to .64 and were significant at .01. In general, by taking a closer look at the model-fit indexes it can be concluded that the model perfectly fits with *RMSEA* = 0.050, $\chi^2/df$=1.69 values (Tabachnick & Fidell, 2001; Jacobucci, Grimm & McArdle, 2016).

As a result of the confirmatory factor analysis, it can be told that the adapted instrument was confirmed to be a valid measurement tool for teacher candidates' computer attitudes. These values indicate that model-data compatibility was sufficient as supported by the literature (e.g., Ingles, Hidalgo & Mendez, 2005; Hoe, 2008). All the sub-factors were consistent with the original sub-factors in the source instrument. Additionally, it can be concluded that the adapted instrument can be used as a valid and reliable measurement tool for determining teachers' computer attitudes. Additionally, by using this instrument more comprehensive intercultural studies can be completed in experimental and action studies.

Also, measuring teachers' attitudes towards computers can contribute to the quality of in-service training about computer and technology for teachers. Specifically when we evaluate teachers' attitudes based on the sub-factors of the adapted instrument, we would know teachers' *interest in, confidence to, adaptation to,* and *perception of* using computers. Accordingly, based on such results the quality of education might be improved. Thus, teachers would be more sensitive in using technology in their educational process and in their daily lives. By offering appropriate education based on computer skill needs in our age, we would have active participants in international platforms. In addition, by using the adapted instrument in different meta-analytic studies would give us feedback in necessary evaluations. Many dimensions, which are absent from the studies in the literature, can be measured with this adapted instrument. As a result, this instrument can be suggested for the use in Turkish academic studies, as a reliable, valid and stronger instrument.

## 6. REFERENCES


Altun, T. (2011). İlköğretim öğrencilerinin bilgisayara yönelik tutumlarının incelenmesi: Trabzon ili örneği. *Turkish Journal of Computer and Mathematics Education, 2*(1), 69-86.

Aşkar, P. & Umay, A. (2001). İlköğretim matematik öğretmenliği öğrencilerinin bilgisayarla ilgili öz-yeterlik algısı. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *21*, 1-8.

Aypay, A. & Özbaşı, D. (2008). Öğretmenlerin bilgisayara karşı tutumlarının incelenmesi. *Kuram ve Uygulamada Eğitim Yönetimi*, *55*, 339-362.

Bahar, E. & Kaya, F. (2013). Meslek yüksekokulu sosyal programlar öğrencilerinin bilgi teknolojileri kullanımlarına yönelik tutumları. *Yükseköğretim ve Bilim Dergisi*, *3*(1), 70-79.

Bahar, H. H., Uludağ, E. & Kaplan, K. (2009). An investigation of the computer and internet attitudes on primary school teachers from Kars province. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi*, *11*(2), 67-83.

Başol, G., & Çevik, V. (2006). *Gaziosmanpaşa Üniversitesi Eğitim Fakültesi öğretim elemanları ve öğrencilerinin bilgisayara yönelik tutumları ile internet kullanım alışkanlıklarının karşılaştırılması* VII. Ulusal Fen ve Matematik Eğitimi Kongresi Kongre Kitabı, *1*, 127-131. Gazi Üniversitesi, Ankara.

Berberoğlu, G. & Çalıkoğlu, G. (1991). Türkçe bilgisayar tutum ölçeğinin yapı geçerliliği. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *24*(2), 841-845.

Bindak, R. & Çelik, H. C. (2006). Öğretmenler İçin Bilgisayar Tutum Ölçeğinin Güvenirlik ve Geçerlik Çalışması. *Eurasian Journal of Educational Research*, *22*, 38-42.

Büyüköztürk, Ş. (2002). *Sosyal Bilimler için Veri Analizi El Kitabı*, Ankara: PegemA Yayıncılık.

Büyüköztürk, Ş., Kılıç-Çakmak E., Akgün, Ö. E., Karadeniz, Ş. & Demirel, F. (2011). *Bilimsel Araştırma Yöntemleri* (10th Ed). Ankara: PegemA Yayıncılık.

Christensen, R. W. & Knezek, G. A. (2009). Construct validity for the teachers' attitudes toward computers questionnaire. *Journal of Computing in Teacher Education*, *25*(4), 143-155.

Cüre, F. & Özdener, N. (2008). Öğretmenlerin bilgi ve iletişim teknolojileri (BİT) uygulama başarıları ve BİT'e yönelik tutumları. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *34*, 41-53.

Çavuş, H. & Gökdaş, İ. (2006). Eğitim Fakültesi'nde öğrenim gören öğrencilerin internetten yararlanma nedenleri ve kazanımları. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 3*(2), 56-78.

Çelik, H. C. & Bindak, R. (2005). İlköğretim okullarında görev yapan öğretmenlerin bilgisayara yönelik tutumlarının çeşitli değişkenlere göre incelenmesi. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 10,* 27- 38.

Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş.(2012). *Sosyal bilimler için çok değişkenli SPSS ve LİSREL uygulamaları*. Ankara: PegemA Yayıncılık

Cohen, P., Cohen, J., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral analysis*. Mahwah, NJ: Erlbaum.

Dorman, J. P., & Knightley, W. M. (2006). Development and validation of an instrument to assess secondary school students' perceptions of assessment tasks. *Educational Studies*, *32*(1), 47-58.

Deniz, L. (2000). Öğretmen adaylarının bilgisayar yaşantıları ve bilgisayar tutumları. *M. Ü. Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi. 12*, 135–166

Deniz, K., Z. (2007). Psikolojik ölçme aracı uyarlama. *Ankara Üniversitesi Eğitim Bilimleri Dergisi*, *40*(1), 1–16

Deniz, L. & Köse, H. (2003). Öğretmen adaylarının bilgisayar yaşantıları ve bilgisayar tutumları arasındaki ilişkiler. *M.Ü. Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi 18*, 39-64.

Demir, Ö. & Yurdugül, H. (2014). Ortaokul ve lise öğrencileri için bilgisayara yönelik tutum ölçeğinin Türkçe'ye uyarlanması. *Eğitim ve Bilim*, *39* (176), 247-256.

DeVellis, R. F. (2003). *Scale development: Theory and applications*. Second Edition. Thousand Oaks: Sage Publications.

Ekici, G , Uzun, N , Sağlam, N . (2010). İlköğretim Öğrencilerinin Bilgisayar Kullanma Sıklığına Bağlı Olarak Bilgisayara Yönelik Tutumlarındaki Değişimin Değerlendirilmesi. İlköğretim Online, 9 (2), 658-667..

Erkan, S. (2004). Öğretmenlerin bilgisayara yönelik tutumları üzerine bir inceleme. *Kırgızistan-Türkiye Manas Üniversitesi Sosyal Bilimler Dergisi*, *12*, 141-145.

Güzeller, C. O. (2011). PISA 2009 Türkiye örnekleminde öğrencilerin bilgisayar öz-yeterlik inançları ve bilgisayar tutumları arasındaki ilişkinin incelenmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, *12*(4), 182-203.

Hashim, H. R. H., & Mustapha, W. N. (2004). Attitudes toward learning about and working with computers of students at UITM. *TOJET: The Turkish Online Journal of Educational Technology*, *3*(2), 3-7.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: technical issues and methods. *Bulletin of the International Test Commission*, *18*, 3-32.

Hambleton, R. K., & Kanjee, A. (1993). Enhancing the validity of cross-cultural studies: improvements in instrument translation methods. *The Annual Meetings of the American Educational Research Association*: Atlanta, GA.

Hung, D. W. L., & Koh, T. S. (2004). A social-cultural view of information technology integration in school contexts. *Educational Technology*, *44*(2), 48-53.

Hoe, S. L. (2008). Issues and procedures in adopting structural equation modeling technique. *Journal of Applied Quantitative Methods*, *3*(1), 76-83.

Kay, R. H. (1990). Predicting student teacher commitment to the use of computers. *Journal of Educational Computing Research*, *6*(3), 299-309.

Kinzie, M. B., & Delcourt, M. A. (1991). Computer technologies in teacher education: The measurement of attitudes and self-efficacy. *The Annual Meeting of the American Educational Research Association*: Chicago, IL.

Knezek, G., Christensen, R., & Miyashita, K. (1998). Instruments for assessing attitudes toward information technology. In Demir, Ö. & Yurdugül, H. (2014). Ortaokul ve lise öğrencileri için bilgisayara yönelik tutum ölçeğinin Türkçe'ye uyarlanması. *Eğitim ve Bilim*, *39* (176), 247-256.

Inglés, C. J., Hidalgo, M. D., & Méndez, F. X. (2005). Interpersonal difficulties in adolescence. *European Journal of Psychological Assessment*, *21*(1), 11-22.

Lehimler, E. (2016). Müzik Öğretmeni Adaylarının Bilgisayar Destekli Öğretime İlişkin Tutum ve Öz-Yeterlik Algılarının İncelenmesi. *Electronic Turkish Studies*, *11*(14).

Levine, T., & Donitsa-Schmidt, S. (1998). Computer use, confidence, attitudes, and knowledge: A causal analysis. *Computers in Human Behavior*, *14*(1), 125-146.

Loyd, B. H. ve Gressard, C. (1984). Reliability and factorial validity of computer attitude scale. In Berberoğlu, G. & Çalıkoğlu, G. (1991). Türkçe bilgisayar tutum ölçeğinin yapı geçerliliği. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *24*(2), 841-845.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555-566.

Khine, M. S. (2001). Attitudes Toward Computers Among Teacher Education Students in Brunei Darussalam. In Erkan, S. (2004). Öğretmenlerin bilgisayara yönelik tutumları üzerine bir inceleme. *Kırgızistan-Türkiye Manas Üniversitesi Sosyal Bilimler Dergisi*, *12*, 141-145.

McInerney, V., McInerney, D. M., & Sinclair, K. E. (1994). Student teachers, computer anxiety and computer experience. *Journal of Educational Computing Research*, *11*(1), 27-50.

McMillan, J. H. (2012). *Educational research: Fundamentals for the consumer* (6th ed.). Boston, MA: Pearson Education Inc.

Mitzner, T. L., Rogers, W. A., Fisk, A. D., Boot, W. R., Charness, N., Czaja, S. J., & Sharit, J. (2016). Predicting older adults' perceptions about a computer system designed for seniors. *Universal Access in the Information Society*, *15*(2), 271-280.

Mumcu, H. Y., & Usta, N. D. (2014). Öğretmen adaylarının bilgisayar ve internet kullanımına yönelik tutumları. *Journal of Instructional Technologies & Teacher Education*, *3*(3), 44-55.

Oğuz, E., Ellez, A. M., Akamca, G. Ö., Kesercioğlu, T. İ., & Girgin, G. (2011). Okulöncesi öğretmen adaylarının bilgisayar destekli eğitim yapmaya ve bilgisayara yönelik tutumları. *İlköğretim Online*, *10*(3), 934-950.

Ozden, M., Aktay, S., Yilmaz, F., & Ozdemir, D. (2007). The Relation between Pre-Service Teachers' Computer Self-Efficacy Believes and Attitudes Towards Internet Use. *International Journal of Learning*, *14*(6), 53-60.

Rana, N. (2012). A study to access teacher educators' attitudes towards technology integration in classrooms. MIER Journal of Educational Studies, Trends & Practices, 2(2), 190–205.

Savaşır, I. (1994). Ölçek uyarlamasındaki sorunlar ve bazı çözüm yolları. *Türk Psikoloji Dergisi*, *9*(33), 27-32.

Slough, S. W., & Chamblee, G. E. (2000). Implementing technology in secondary science and mathematics classrooms: A perspective on change. *Society for Information Technology & Teacher Education International Conference*. *1*, 1021-1026.

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar. *Türk Psikoloji Yazıları*, *3*(6), 49-74.

Şahin, A., & Akçay, A. (2011). Türkçe öğretmeni adaylarının bilgisayar destekli eğitime ilişkin tutumlarının incelenmesi. *Electronic Turkish Studies*, *6*(2), 909-918.

Şimşek, Ö.F. (2007). *Yapısal Eşitlik Modellemesine Giriş: Temel İlkeler ve LISREL Uygulamaları*. Ankara: Ekinoks Yayınları

Tabachnick, B. G. & Fidell, L. S. (2001). *Using multivariate statistics*, Fourth Edition. New York: Harper Collins Publishers.

Tavşancıl, E. (2014). *Tutumların ölçülmesi ve SPSS ile veri analizi*. 5. Baskı. Nobel Yayın Dağıtım.

Teo, T. (2008). Assessing the computer attitudes of students: An Asian perspective. *Computers in Human Behavior*, *24*, 1634-1642.

Teo, T. (2009). The impact of subjective norm and facilitating conditions on pre-service teachers' attitude toward computer use: A structural equation modeling of an extended technology acceptance model. *Journal of Educational Computing Research*, *40*(1), 89-109.

Teo, T., Milutinović, V., & Zhou, M. (2016). Modelling Serbian pre-service teachers' attitudes towards computer use: A SEM and MIMIC approach. *Computers & Education*, *94*, 77-88.

Usta, E., & Korkmaz, Ö. (2010). Öğretmen adaylarının bilgisayar yeterlikleri ve teknoloji kullanımına ilişkin algıları ile öğretmenlik mesleğine yönelik tutumları. *Uluslararası İnsan Bilimleri Dergisi*, *7*(1), 1335-1349.

Yeşilyurt, S., & Gül, Ş. (2007). Bilgisayar kullanma becerileri ve bilgisayarlara yönelik tutum ölçeği (BKBBYTÖ): Geçerlik ve Güvenirlik Çalışması. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi*, *24*, 79-88.

Yıldırım, S., & Kaban, A. (2010). Öğretmen adaylarının bilgisayar destekli eğitime karşı tutumları. *Uluslararası İnsan Bilimleri Dergisi*, *7*(2), 158-168.

*Int. J. Asst. Tools in Educ., Vol. 4, Issue 2, (2017) pp. 182-210*

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

*"Research Article"*

# Parental Perceptions about Children's Authentic Assessment and the Work Sampling System's implementation

**Anastasios Pekis**[1,*] **Efthymia Gourgiotou**[1]

[1]Department of Preschool Education, Faculty of Education, University of Crete, Greece

***Abstract***

The authentic assessment process in preschool education gains the confidence of the practices which are used today, as an innovative educational policy in the interest of everyone who's involved in early childhood education: children, teachers and parents. The purpose of this study was to explore parents' perceptions upon the significance of child's assessment, their engagement in this assessment and the impact of the implementation of alternative forms of assessment such as the Work Sampling System at the kindergarten. A survey research design was utilized in order to achieve the objectives of the study, where a small-scale questionnaire was given to a convenience sample of 18 parents whose children were enrolled in a public all-day kindergarten in Chania, Greece. Findings show that the majority of the parents either acknowledge children's authentic assessment as a real breakthrough or they are satisfied on a large scale with the implementation of alternative forms of assessment in the classroom. In conclusion, the child's authentic assessment has been recognized widely as it is advantageous to the educational settings of the modern pedagogy.

***Article Info***

***Received***
*28 February 2017*
***Revised***
*28 May 2017*
***Accepted***
*31 May 2017*

***Keywords***
Authentic assessment;
parents' perceptions;
preschool education;
Work Sampling System

## 1. INTRODUCTION

In most early childhood programs, early childhood educators use a variety of kindergarten assessment instruments aiming to give an accurate picture of children's development and learning throughout the school year. Decades of research on the assessment of the child have evidenced that alternative forms of assessment are the most powerful tools as authentic exhibits of improved developmental pathways and learning outcomes of preschool children in all areas of learning

[1] Corresponding Authors' Email: apekis@edc.uoc.gr egourgiotou@edc.uoc.gr

2148-7456 /© 2017 DOI: 10.21449/ijate.318250

suggested by the kindergarten curriculum (Gullo, 2005; Bagnato, 2007; Snow & Van Hemel, 2008; Losardo & Notari-Syverson, 2011; Fiore, 2012; Frey, Schmitt, & Allen, 2012).

In the early years, as research has shown, authentic measures are emphasized more than traditional forms (Bergen,1993; Grisham-Brown, Hallam, & Brookshire, 2006; Bagnato, 2007). According to numerous research studies, authentic assessment is deemed a significant tool of the teaching and learning process. In the light of pedagogical science, authentic assessment can be defined as a systematic procedure of collecting and analyzing important information and evidence that teachers use to understand holistically children's progress in all domains of development in natural classroom contexts (Henderson & Karr-Kidwell, 1998; Wortham, 2008; Losardo & Notari-Syverson, 2011; Swaffield, 2011). Authentic assessment can include some of the following: teacher observations and records, portfolios, rubrics, self and peer assessments, performance-based assessment, naturalistic assessment, play-based assessment (Gullo, 2005; Doliopoulou & Gourgiotou, 2008; Brodie, 2013). With respect to evaluation methods which are used in education, authentic assessment is more appropriate than traditional assessment in the kindergarten because it reflects children's learning and achievement on classroom activities taking into account the significance of real-life contexts and the natural learning environment of the child in the preschool setting.

The assessment of young children in preschool environment, according to several studies, contains three important and specific elements: (a) documentation process, (b) evaluation, and (c) partnership and communication with children's parents (Johnson, 1993; Hannon, 1997; Carr, 2001; Lam, 2008). Acknowledging the fact that assessment is an ongoing procedure, the use of different methods of documentation constitutes a concrete way of tracking children's progress in all domains of learning. Additionally, applying assessment strategies that are developmentally appropriate and child-centered for preschoolers is undeniably the key to significant positive ramifications and changes on students' performance and on teachers' instructional and learning strategies (Shepard, 1994; Brookhart, 2004; Wortham, 2008; Copple & Bredekamp, 2009). Also, bridging the potential gap between parents and school, and engaging parents as partners in children's education can become effective in tutoring and in facilitating each child's growth, development and acquisition of knowledge since home-school collaboration can give significant information to both enmeshed sides (Work & Stafford, 1987; Gelfer, 1991; Billman, Geddes, & Hedges, 2005; Peters, Seeds, Goldstein, & Coleman, 2008).

According to the research, the personal school experiences and the bias of parents affect their perceptions about assessment methods in the school community. Quite a few parents are suspicious and show hesitancy towards authentic assessments (Shepard & Bliem, 1995). Understanding the parents' perceptions about children's assessment is an important issue for a number of reasons. These reasons include: (a) the misconceptions among parents about assessment in kindergarten or the lack of education of what child's assessment refers to, (b) to provide valuable insights into design of the assessment measures used in the kindergarten or program quality improvement plans, (c) to give multiple valid perspectives to parents that will inform them about the quality and the significance of the children's assessment, (d) to increase parents' understanding of the appropriate assessment practices used in the context of the preschool setting and the reasons they are implemented, (e) to enhance teacher's instructional practices and decisions for children's benefit and, (f) to involve parents and teachers in a collaborative context that will support and promote the child's development and will make children's thinking and learning visible.

To meet the appropriate standards for a successful assessment and try to acquire a balance among the above-mentioned factors, it is important to discern parents' views and convictions on child's assessment as parenthood is considered crucial at this stage of child development. A number of authors have pointed out that parents should be provided with teachers' evaluations on children's progress with profound updates, involved in school conferences and considered as a valued source of assessment information (Shepard & Bliem, 1995; Culbertson & Jalongo, 1999; Finello, 2011; Orillosa & Magno, 2013; Birbili & Tzioga, 2014). Early childhood practitioners and parents have the right to be conversant with the strengths and needs of children in order to provide effective support and learning opportunities either in the school setting or within the family environment (Brink, 2002; Hill & Taylor, 2004; Clinton & Guilar, 2016).

Taking into account the significance of children's assessment in kindergarten, an attempt is made by the present study to explore and look into parental perceptions about: (a) the children's assessment in the kindergarten in general, (b) their engagement in children's assessment and (c) the impact of the implementation of WSS, as an authentic assessment tool, in particular in the following parts.

### 1.1. The Challenge of Supporting Authentic Assessment in Preschool Education

The issue of authentic assessment in kindergarten has been identified by the researchers to a considerable extent as a significant procedure used for varied purposes. When referring to kindergarten community, assessment in the first school years is essential as it consists a key component to understand children's development in the early years. Taking into account that previous studies acknowledge the importance of parental involvement in children's learning (Hill & Taylor, 2004; Galindo & Sheldon, 2012), authentic assessment constitutes the appropriate context for the stakeholders to collaborate. Indeed, this type of evaluation involves children, educators and parents in an active way and promotes positive outcomes for everyone (Brink, 2002; Palm, 2008; Swaffield, 2011). In particular, authentic assessment is referred to as a systematic approach that collects data and useful information from children, teachers and parents reflecting and emphasizing on children's learning, achievement, real-life competencies in everyday routines over time and in real conditions (Hart, 1994; Bagnato, 2007; Doliopoulou & Gourgiotou, 2008; Riley, Miller, & Sorenson, 2016). Getting to the heart of authentic assessment, the literature highlights the importance of using alternative forms of assessment in any educational procedure (Dennis, Rueter, & Simpson, 2013). Authentic assessment approach recognizes the active role children play in acquisition of knowledge in natural settings or in pointed realistic tasks (Brassard & Boehm, 2007).

Assessment practices may be implemented through the use of various techniques and strategies that can be adapted for different situations in order to track children's progress in all areas of learning. According to Losardo and Notari-Syverson (2011), gaining insights into children's learning, needs, strengths and interests can be accomplished by observing children and documenting their work, considering them as the most common and appropriate ways in the context of children's evaluation. In the above context of this alternative assessment method, evaluation of the child is a shared responsibility of those who are involved in the educational process. In the authentic assessment environment, teachers and children can act effectively in the school community and set targets to improve the quality of teaching and learning process. Educators need to combine authentic assessment techniques with daily practice interpreting assessment as a part of effective planning of teaching and learning and not as an isolated event in

the daily school routine (Darling-Hammond & Snyder, 2000; Downs & Strand, 2006; Bagnato, 2007; Wortham, 2008).

The challenge of supporting and utilizing alternative assessment approaches in early childhood education can contribute positively to teaching and learning. What research studies have shown over the last three decades is that authentic assessment constitutes an integral element of educational practice and is deemed necessary in order to: (a) specify the children's strengths, interests and needs, (b) identify and document children's achievement over time, (c) diagnose children who may be in need of specialized training, (d) support each child's self-confidence and self-esteem, (e) help children comprehend their personal learning advancement through critical thinking, reflection and feedback, (f) aid towards making appropriate instructional decisions or future instructions suited to the context of classroom (g) improve the educational program and its desired outcomes in a qualitative way and (h) give information to parents or other teachers of primary education (Epstein, Schweinhart, Debruin, & Robin, 2004; Grisham-Brown et al., 2006; Doliopoulou & Gourgiotou, 2008; Bagnato, McLean, Macy, & Neisworth, 2011; Dennis, Rueter, & Simpson, 2013).

As described earlier, it is clear and quite obvious that authentic assessment serves plenty of pedagogical purposes in the context of early childhood education as it is considered essential by policy makers, teachers, children and parents.

### 1.2. What Parents Know About Children's Assessment?

Another key feature of authenticity relevant to early childhood assessment is communication with family. Family involvement in preschool education can strengthen and support to a great extent children's well-being in social, cognitive and emotional level in a variety of appropriate ways. There is clear evidence that early childhood educational programs, curriculum standards, policies, school community, taking into account and responding effectively to the learning needs of all children encourage and emphasize strongly on building collaboration and partnership programs among parents and educators (Work & Stafford, 1987; Billman et al., 2005; Doliopoulou & Gourgiotou, 2008; Murray, Curran, & Zellers, 2008). Many aspects of effective authentic assessment require collaboration with families and kindergarten teachers. Parents have the right to be informed about how their children are doing in kindergarten and get an accurate picture of their school learning and improvement (Engel, 1993; Olmscheid, 1999). By showing simple examples of the daily kindergarten routine to parents, they are enabled to personally assess their children's growth and progress. Since parents have the right to access information about children's progress, this fact itself is a principal characteristic of education policies that give value to the practices which facilitate and promote authentic assessment tools in preschool practice (Dafermou, Koulouri, & Basagianni, 2006; NAEYC, 2009; Hall, Rutland, & Grisham-Brown, 2011).

In the light of the survey findings, children's learning and personal development constitute a shared responsibility for both teachers and parents (Becher, 1984; Baum & McMurray-Schwarz, 2004). It is particularly important to take into account parents' views on kindergarten assessment practices because they are considered as a significant factor in the whole school system. Research background indicates that parental perceptions about children's authentic assessment is an important issue that has been an ongoing concern for the researchers over the last decades but unfortunately the majority of these studies mainly sampled primary school parents and not kindergarten parents that often. Most parents, as data research indicates, support the use of authentic assessment in kindergarten (Shepard & Bliem, 1995; Hannon, 1997; Culbertson &

Jalongo, 1999; Osburn, Stegman, Suitt, & Ritter, 2004). Patricia Atkinson (2003) highlighted in her action research the importance of classroom assessment and the parental reports concerning useful information about the child's progress and not just summative types of assessment. Talking with families about children's assessment is a positive way to establish constructive home-school interactions, relationships or information exchange for the benefit of all children.

The trend to use alternative approaches of assessment and reporting is supported strongly in Meisels, Xue, Bickel, Nicholson, and Atkins-Burnett's (2010) study. The forenamed researchers have found that parents are supportive to performance assessment under the two following circumstances: (a) when school communities use systematically these assessments and (b) when school implements consistent informal communications between parents and educators.

In conclusion, this short literature review indicated that over the last three decades there has been an important change in assessment in early childhood education moving from formal testing to alternative forms of assessment.

### 1.3. The Structure of Work Sampling System: A General Overview

Work Sampling System (WSS) constitutes an instructional assessment tool that uses: (a) guidelines and checklists: a set of observational criteria to assist teachers focus on observation and evaluate student performance, (b) portfolios: unique collections of children's work and progress, and (c) summary reports: written informational reports on student performance and progress based on teachers' observations and documentation, checklist ratings and portfolio work (Dichtelmiller, Jablon, Dorfman, Marsden, & Meisels, 2001).

WSS contributes to monitoring children's self-growth by teachers across seven developmental domains: personal and social development, language and literacy, mathematical thinking, scientific thinking, social studies, the arts and, physical development, health, and safety. Teachers make ratings three times per year at the end of each data collection period (autumn, winter, spring) using WSS Developmental Guidelines, creating in this way the profile of children's personal progress and the real duties they have to perform in different developmental areas. The process of collecting information systematically on what children have done or learned, and the evaluation of this information constitute two significant steps for WSS that teachers must follow when applying it in the classroom (Meisels, 1993).

The purpose of these three elements of WSS is to help educators document and assess children's academic skills, learning level, behaviors and school performance during their schooling from kindergarten to primary school in an appropriate way (Meisels, 2011). The worthiness of WSS is based on its use as an innovative systematic approach of children's learning progress during the school year. It is mainly based on the compilation of children's work and teachers' observations and documents collected from everyday experiences, routines, free and organized activities implemented in an authentic learning environment. It involves children, teachers and parents in the learning and assessment procedure, providing and sustaining meaningful feedback for the stakeholders (Meisels, 1997).

### 1.4. A Brief Critical Review of Work Sampling System

As it is mentioned above, Work Sampling System (WSS) emphasizes on the teacher's observations and on the processes that children utilize in order to acquire knowledge through authentic situations such as the classroom setting. According to Meisels (1997), the plurality of data and information emerged using teacher's observations, portfolios, developmental checklists

and summary reports, strengthens the learning and teaching process and outlines in detail each student's profile.

Meisels, Liaw, Dorfman and Nelson (1995) emphasize that the WSS implementation, as an alternative assessment tool, is a reliable and valid approach for assessing the learning progress of kindergarten children. In the study mentioned above, findings show that WSS can yield valid and adequate results as compared to traditional forms of assessment. Subsequent surveys with a larger sample of children ranged in age from 5 years to 10 years, confirmed and expanded the previous findings concerning the reliability, the validity and the consistency of teacher observations through the WSS implementation (Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001; Meisels et al., 2010).

Also, the parental involvement in the children's assessment is an important aspect which is directly linked to the WSS philosophy. The mentioned assessment tool improves the cooperation among teachers and parents and fosters family involvement in the educational process. On account of this, teachers' meetings with parents at school are considered essential, as they are informed about children's performance and progress throughout the school year. Relevant findings are presented by the study of Meisels et al. (2001), in which parents have a positive attitude towards the information they receive and the benefits of the WSS implementation to their children.

The WSS is not offered free of charge, as it consists a commercial product available in paper and online. Nevertheless, the current research acknowledges the importance of children's assessment in preschool education by presenting the WSS assessment tool as an example of alternative practices in the assessment of young children. Public kindergartens and preschool educators could implement an authentic assessment based on the structure and the principles of the WSS assessment tool. It is recommended that kindergarten teachers find authentic forms of assessing their children and adapt or design an appropriate assessment tool, keeping in mind the principles and the purposes of assessment, the adequacy of the assessment techniques, the learning styles of each child and the inclusion of families.

## 2. METHOD

The concept of children's assessment in kindergarten nowadays is considered to be a significant issue that has been of great concern to the educational community. Many researchers reveal the value of children's assessment from preschool years and recognize its importance generally to the educational process in the classroom (Appl, 2000; Epstein et al., 2004; Sakellariou, 2006; Doliopoulou & Gourgiotou, 2008; Kazela & Kakana, 2009). The absence of an identifiable systematic research in Greece on this thematic unit directed the researchers to the survey as a first attempt to map the current situation. The present research was deemed essential as a part of gaining an understanding of the needs and concerns of parents with regard to children's assessment in Greek preschool education. The significance of early childhood assessment, the parents' perspectives in early childhood assessment and the implementation of innovative assessment practices play a vital role in the early years as they constitute co-depended parameters of the educational process.

Based on the needs of relevant literature, the purpose of this study is to investigate parents' perceptions towards the role and the function of the child's authentic assessment in preschool education, the impact of the application of alternative forms of assessment such as the Work Sampling System implementation and how parents respond to and perceive those evaluation

methods. The understanding and knowledge gained from this study will benefit teachers, families and communities, advancing educational practices and policies in the context of early childhood education.

### 2.1. Research Questions

In order to specify the parental perceptions on child's assessment in kindergarten, the following research questions guided this study:

1. What are parent's perceptions of the assessment implementation and its significance in the kindergarten classroom?
2. How do parents react to WSS—the performance assessment in use—overall?
3. Which specific factors affect parents' overall perceptions to WSS?
4. To what extent is the role of the implementation of the child's evaluation in preschool education related to the role of the portfolio, the children's developmental checklists and the kindergarten teacher's summary reports?
5. What are parents' perceptions about their engagement in children's assessment?

### 2.2. Study design

A research survey was designed and implemented by the researchers in order to explore and answer the research questions. While conducting a research survey it becomes clear that there are many benefits such as reliability and flexibility, high representativeness, low cost, convenient data gathering (Cohen, Manion, & Morrison, 2008; Tuckman & Harper, 2012). Additionally, the research followed the principles of case study, as it was implemented in a public all-day kindergarten. The design of the present study is quantitative and it was conducted as an exploratory case study.

There are several categories of case study. Yin (1994) notes three categories, namely exploratory, descriptive and explanatory case studies. First, exploratory case studies set to explore any phenomenon in the data which serves as a point of interest to the researcher. For instance, a researcher conducting an exploratory case study on individual's reading process may ask general questions, such as, "Does a student use any strategies when he reads a text?" and "if so, how often?". These general questions are meant to open up the door for further examination of the phenomenon observed. In this case study, also, prior fieldwork and small-scale data collection may be conducted before the research questions and hypotheses are proposed. As a prelude, this initial work helps prepare a framework of the study. A pilot study is considered an example of an exploratory case study (Yin, 1984; McDonough & McDonough, 1997) and is crucial in determining the protocol that will be used. By using the case study method, researchers can explore, examine and explicate data in real-life context. This type of approach can be exploratory, constructive or confirmatory when there is a need to obtain an in-depth appreciation of an issue, event or phenomenon of interest (Creswell, 2012; Bryman, 2016).

### 2.3. Research Process

The study was conducted during the school year 2015-2016 in a public all-day kindergarten in the prefecture of Chania in Greece by the kindergarten teacher himself. The parents whose children were enrolled in this kindergarten participated in this study (n=18). The WSS test was translated in Greek according to the developmental directions of the Greek Cross-Thematic Curriculum Framework and the Kindergarten Teachers' Guide. Permission was given by the authors to translate and use the assessment tools. The translation process was carried out according

to the study published by Beaton, Bombardier, Guillemin and Ferraz (2000) as follows: (i) Translation into Greek: the WSS was translated by two native-speaking Greek translators, resulting in translations T1 and T2. (ii) Analysis: both translations were analyzed to reach a consensus on a single translation (T12). (iii) Back translation to English: native-speaking English translators, who were unaware of the process carried out, translated T12 into two new English versions (RT1 and RT2). (iv) Revision by an expert committee: the committee was comprised of four occupational therapists, two translators, and two researchers familiar with the tool. After analyzing all the versions (T1, T2, T12, RT1, and RT2), a pre-final Greek version was chosen. (v) Pilot testing of the pre-final version: in this phase, the pre-final version was used on a sample of 20 children, 18 parents and two kindergarten teachers, who were subsequently interviewed regarding any difficulties they had in understanding the meaning of the questions and the responses. Incidents of non-completed and repeated replies were also analyzed (i.e., when all participants provide the same response to a specific question). (vi) Use of the tool: the present pilot WSS was designed and the tool was administered to 40 people, including two teachers, 20 children and 18 parents. (vii) Conclusions: errors and typing mistakes which derived in the final version of the WSS were checked. This version was then sent to the authors of the original WSS.

At the start of the school year 2015-2016, parents were informed in detail about the assessment tool, in order to create a positive framework for cooperation and to point out the significance of the use of authentic assessment in kindergarten.

In the first phase, the kindergarten teachers informed the parents generally about the use of alternative forms of assessment in kindergarten and presented, in particular, the WSS assessment tool and its components. Instructions about its use were given and clarification questions were answered in order to highlight the effectiveness of a reliable and valid assessment tool in preschool children such as the Work Sampling System (WSS). After presentation, the parents signed the consent form for participation in the research, according to the instructions of the Greek Institute of Educational Policy (IEP).

In the second phase, the kindergarten teachers informed the children about the use of WSS in classroom. Each child had his own folder which included: (a) the WSS checklist, (b) a portfolio folder, and (c) Summary Reports of kindergarten teachers about child's development. Near the end of the first term, teachers used the WSS Developmental Guidelines book to rate children based on their observations and the documents in the children's portfolios. Completed checklists and summary reports were announced in meetings with parents three times per year in order to provide useful information about children's performance, skills, knowledge and behaviors. Checklists and summary reports were also used in order to plan developmentally appropriate classroom experiences throughout the school year by the teachers. At the end of each term, each family kept the WSS evaluation tool at home for a week helping the family feedback and reflected on child's achievements.

At the end of the school year 2015-2016, the kindergarten teachers organized a meeting with parents and discussed the benefits of children's assessment in kindergarten. The kindergarten teachers and parents discussed their aspirations and the center philosophy of children's assessment together. All participants were asked to look back over the year at their children's progress in order to share understanding and knowledge about children's assessment, their perceptions and final reports about the implementation of WSS during the school year and their views about future goals. Parents had a meaningful and productive discussion as they expressed their viewpoints and concerns realizing the positive outcomes of assessment at the preschool setting.

## 2.4. Sample - Participants

The sample of the research consisted of 18 parents, whose children were enrolled in a public all-day kindergarten in the prefecture of Chania during the school year 2015-2016. The demographic characteristics of the sample are described in Table 1.

**Table 1.** Demographic characteristics of the sample

| | | Frequency | Percentage |
|---|---|---|---|
| Gender | Male | 5 | 27.8 |
| | Female | 13 | 72.2 |
| SUM | | 18 | 100.0 |
| Age | 23-33 | 5 | 27.8 |
| | 34-44 | 13 | 72.2 |
| Employment status | Civil servant | 5 | 27.8 |
| | Private employee | 8 | 44.4 |
| | Self-employed | 4 | 22.2 |
| | Else | 1 | 5.6 |
| Educational level | Compulsory Secondary Education | 1 | 5.6 |
| | Post-Compulsory Secondary Education | 12 | 66.7 |
| | University education/Technical Educational Institute | 5 | 27.8 |
| Household composition | Two people | 2 | 11.1 |
| | Three people | 2 | 11.1 |
| | Four people | 9 | 50.0 |
| | Five people | 4 | 22.2 |
| | More than five people | 1 | 5.6 |
| Marital status | Married | 17 | 94.4 |
| | Divorced | 1 | 5.6 |

Specifically, it seems that 94.4% of the sample are married and females form the majority of the sample. Furthermore, the majority of the sample belongs to the age group of 34-44 and as far as the employment status of parents is concerned, 44.4% are private-employees and 27.8% are civil servants. Moreover, 66.7% of the sample are graduates of the Post-Compulsory Secondary Education and 27.8% of the sample are graduates of the University Education. Furthermore, half of the participants (50.0%) said that their household consisted of four people.

## 2.5. Data Collection Tool

Data collection in this study consisted of one questionnaire which was developed by the researchers and was divided into three parts. The first part comprised closed-ended questions about the demographic characteristics of the sample. The second part consisted of a closed-ended question about the significance of the implementation of children's assessment in preschool education. The third part comprised: (a) closed-ended questions about the importance of portfolio assessment, developmental checklists and summary reports of WSS in the kindergarten, (b) open-

ended questions regarding the benefits and the drawbacks of the WSS application in kindergarten during the school year, the presence of parental involvement or not in the children's evaluation process and additional opinions relevant to the child's assessment. The questions of the second part were designed on a 5-point Likert scale (not at all, a little, enough, a lot, very much) and the questions of the third part on a 5-semantic differential scale (1=minimum, 2, 3, 4, 5=maximum). All the questionnaires were accompanied by a letter explaining the purpose of the research study, ensuring the participants' anonymity and the non-disclosure of personal data.

In the present study, internal consistency of the questionnaire was calculated by Cronbach's Alpha, as the most important and common measure of scale reliability (Field, 2009). The following table showed that the three scales have high internal consistency (0.923) with a range between 0.728 and 0.923, indicating that the researchers' instrument has a good degree of reliability and confirming its use for data collection.

**Table 2.** Reliability analysis of measurement scales (Cronbach's Alpha)

| Scale | Cronbach's Alpha | N of Items |
|---|---|---|
| The role of the implementation of the children's evaluation in preschool education | 0.845 | 7 |
| The role of the portfolio as an assessment tool in kindergarten | 0.728 | 6 |
| The role of the children's developmental checklists and kindergarten teachers' summary reports | 0.923 | 7 |

### 2.6. Data Analysis

Data analyses included: (a) a descriptive analysis to calculate the median, range, frequencies, percentages of parental views, (b) a reliability analysis to examine the reliability of a part of the questionnaire, and (c) a Spearman Rank Correlation to measure relationships. After the surveys were returned, data were encoded and responses were registered on the computer for statistical analysis. The data analysis was performed by using SPSS 21.0, statistical software for Windows.

## 3. FINDINGS

In the first part of the questionnaire reference was made to the demographic characteristics of the research sample. The second part of the questionnaire included a question concerning the parents' views about the significance of the implementation of children's assessment in preschool education. Table 3 presents the level of agreement of the participants regarding the implementation of the children's assessment in preschool education.

**Table 3.** The role of the implementation of the child's evaluation in preschool education

| Assessment in preschool education | Not at all | A little | Enough | A lot | Very much | Median | Range |
|---|---|---|---|---|---|---|---|
| Assessment helps the kindergarten teacher to understand the level of knowledge and skills gained by children. | 0 (0.0) | 0 (0.0) | 1 (5.6) | 2 (11.1) | 15 (83.3) | 5.0 | 2.0 |
| Assessment in kindergarten helps the teacher make instructional design decisions. | 0 (0.0) | 0 (0.0) | 1 (5.6) | 6 (33.3) | 11 (61.1) | 5.0 | 2.0 |
| Assessment enables the kindergarten teacher to assess the performance and the progress of young children. | 0 (0.0) | 0 (0.0) | 1 (5.6) | 1 (5.6) | 16 (88.9) | 5.0 | 2.0 |
| Assessment in kindergarten assists to record the children's learning development during the school year. | 0 (0.0) | 0 (0.0) | 0 (0.0) | 2 (11.1) | 16 (88.9) | 5.0 | 1.0 |
| Assessment in kindergarten facilitates the actual learning of young children. | 0 (0.0) | 0 (0.0) | 2 (11.1) | 2 (11.1) | 14 (77.8) | 5.0 | 2.0 |
| Assessment aids the teacher to identify children with learning difficulties or behavioral problems. | 0 (0.0) | 0 (0.0) | 2 (11.1) | 2 (11.1) | 14 (77.8) | 5.0 | 2.0 |
| Assessment in kindergarten facilitates briefing of the family. | 0 (0.0) | 0 (0.0) | 1 (5.6) | 1 (5.6) | 16 (88.9) | 5.0 | 2.0 |

In the third part of the questionnaire, two questions were included about the importance of portfolio assessment, developmental checklists and summary reports of WSS in the kindergarten and three questions regarding the benefits and the drawbacks of the WSS application in kindergarten, the presence of parental involvement in the children's assessment and additional opinions relevant to the child's assessment. Table 4 presents the percentage of parents' ratings regarding the role of the children's portfolio as an assessment tool. Also, information is provided about the median and the range of their viewpoints.

**Table 4.** The role of the portfolio as an assessment tool in kindergarten

| Portfolio assessment | Lower degree 1 | 2 | 3 | 4 | Higher degree 5 | Median | Range |
|---|---|---|---|---|---|---|---|
| Helps the children to be involved actively in daily kindergarten learning procedures. | 0 (0.0) | 0 (0.0) | 0 (0.0) | 4 (22.2) | 14 (77.8) | 5.0 | 1.0 |
| Helps the children to self-assessment procedure and observe their progress. | 0 (0.0) | 0 (0.0) | 2 (11.1) | 5 (27.8) | 11 (61.1) | 5.0 | 2.0 |
| Helps the children to rethink and reflect on how they did their work or how they acquired knowledge. | 0 (0.0) | 0 (0.0) | 2 (11.1) | 2 (11.1) | 14 (77.8) | 5.0 | 2.0 |
| Helps the children to develop feelings of autonomy, self-esteem, individual choices and pride. | 0 (0.0) | 0 (0.0) | 2 (11.1) | 4 (22.2) | 12 (66.7) | 5.0 | 2.0 |
| Urges the children to express their personal interests, needs and abilities. | 0 (0.0) | 0 (0.0) | 0 (0.0) | 7 (38.9) | 11 (61.1) | 5.0 | 1.0 |
| Helps the children, the kindergarten teachers and the parents to assess potential and possible weaknesses. | 0 (0.0) | 0 (0.0) | 1 (5.6) | 3 (16.7) | 14 (77.8) | 5.0 | 2.0 |

As Table 4 shows, the majority of the parents (77.8 %) seems to perceive the significance of the portfolio as it helps the children largely to be involved actively in daily kindergarten learning procedures. At the same time, the view that portfolio helps the children to rethink and reflect on how they did their work or how they acquired knowledge is supported by the 77.8 % of the sample. Ultimately, in a few cases the portfolio assessment is motivational as it urges children to express their personal interests, needs and abilities.

The value of the use of children's checklists and kindergarten teachers' summary reports is the upcoming research question. Table 5 presents the median and the range of parental views regarding the value of the use of children's developmental checklists and kindergarten teachers' summary reports.

**Table 5.** The value of the use of developmental checklists and summary reports

| Children's developmental checklists and kindergarten teachers' summary reports help parents to understand | Lower degree 1 | 2 | 3 | 4 | Higher degree 5 | Median | Range |
|---|---|---|---|---|---|---|---|
| The way children think and develop | 0 (0.0) | 1 (5.6) | 1 (5.6) | 2 (11.1) | 14 (77.8) | 5.0 | 3.0 |
| The learning process of each child individually in every period of the school year | 0 (0.0) | 0 (0.0) | 1 (5.6) | 5 (27.8) | 12 (66.7) | 5.0 | 2.0 |
| Children's potential weaknesses | 0 (0.0) | 0 (0.0) | 2 (11.1) | 4 (22.2) | 12 (66.7) | 5.0 | 2.0 |
| Children's progress in accordance with the principles and objectives of the kindergarten curriculum | 0 (0.0) | 0 (0.0) | 1 (5.6) | 5 (27.8) | 12 (66.7) | 5.0 | 2.0 |
| The level of knowledge, skills or attitudes children have acquired | 0 (0.0) | 0 (0.0) | 1 (5.6) | 5 (27.8) | 12 (66.7) | 5.0 | 2.0 |
| The potential behavioral problems or learning difficulties of each child. | 0 (0.0) | 0 (0.0) | 3 (16.7) | 3 (16.7) | 12 (66.7) | 5.0 | 2.0 |
| The kindergarten daily program and the cognitive learning areas. | 0 (0.0) | 1 (5.6) | 1 (5.6) | 4 (22.2) | 12 (66.7) | 5.0 | 3.0 |
| Percentage (%) | | | | | | | |

According to Table 5, it seems that the use of children's developmental checklists and kindergarten teachers' summary reports offers an important advantage to parents. They gain an understanding of the multiple ways their children think and develop. Finally, in many instances it is evident that children's developmental checklists and kindergarten teachers' summary reports provide parents with considerable information regarding the learning progress, the potential weaknesses of their children and the function of the kindergarten in relation with the principles of the curriculum.

The following research question concerns the benefits and the drawbacks of the WSS application in kindergarten during the school year. Table 6 shows the frequencies and the percentages of parents' ratings concerning the benefits of the WSS assessment tool.

**Table 6.** The benefits of the use of Work Sampling System

| | Frequency | Percent |
|---|---|---|
| It is an integrated recording of the child's progress and development, according to the kindergarten curriculum. | 6 | 33.3 |
| Children are actively involved in a continuous procedure of development and the learning process is enhanced. | 1 | 5.6 |
| It is an integrated recording of the child's progress and development according to the kindergarten curriculum, an understanding of their potential, weaknesses and knowledge level and a diagnostic means of possible learning-behavioral problems. | 2 | 11.1 |
| It is an understanding of the children's knowledge level, a frequent parental briefing and an indicator of the children's active involvement in the learning process. | 1 | 5.6 |
| It is an understanding of the children's knowledge level and a means of assisting the child's self-assessment. | 1 | 5.6 |
| It is an integrated recording of the child's progress and development according to the kindergarten curriculum and an understanding of the children's potential, weaknesses and knowledge level by teachers and parents. | 3 | 16.7 |
| It is an integrated recording of the child's progress and development according to the kindergarten curriculum, an understanding of the children's knowledge level, an indicator of the active involvement of the children in a continuous procedure of development and an indicator of the enhancement of the learning process. | 1 | 5.6 |
| It is an integrated recording of the child's progress and development according to the kindergarten curriculum, an understanding of the children's potential, weaknesses and knowledge level by teachers and parents, an indicator of the active involvement of the children in a continuous procedure of development and an indicator of the enhancement of the learning process. | 1 | 5.6 |
| It is an integrated recording of the child's progress and development according to the kindergarten curriculum and a pedagogical documentation of children's learning experiences. | 1 | 5.6 |
| It is an integrated recording of the child's progress and development according to the kindergarten curriculum, an understanding of the children's knowledge level, an understanding of the children's potential, weaknesses and knowledge level by teachers and parents and a means that encourages children to express their needs, interests and efforts. | 1 | 5.6 |
| Total | 18 | 100.0 |

As it can be seen in Table 6, it is clear that the most important advantage of WSS implementation in preschool classroom is that the assessment tool is considered as an integral recording of child's progress and development, according to the basic principles of the kindergarten curriculum. Meanwhile, an average percentage of the respondents (16.7%) consider the use of WSS significant because they gain a better understanding of their children's potential, weaknesses and knowledge level both by teachers and parents.

In addition, Table 7 shows the frequencies and the percentages of parents' ratings concerning the drawbacks of the WSS assessment tool.

**Table 7.** The drawbacks of the use of Work Sampling System

| | Frequency | Percent |
|---|---|---|
| As an autonomous assessment tool (WSS) cannot function well, unless kindergarten teachers organize briefings with parents simultaneously. | 1 | 5.6 |
| The incorrect reading and interpretation of WSS by parents can cause real angst or create high expectations from the children. | 1 | 5.6 |
| The absence of a numerical scale does not always help the interpretation and understanding of the child's progress. | 1 | 5.6 |
| It's a time-consuming process for the kindergarten teacher to collect, analyze and interpret the related data concerning the assessment of each child. | 4 | 22.2 |
| There is a possibility of failing to record everything which takes place in the classroom by the teachers. | 2 | 11.1 |
| As an autonomous assessment tool (WSS) cannot function well, unless kindergarten teachers organize briefings with parents simultaneously, it is necessary all three components of WSS be used in parallel otherwise the WSS assessment tool will not be realized to its full extent. | 1 | 5.6 |
| No disadvantages found. | 8 | 44.4 |
| Total | 18 | 100.0 |

In particular, the majority of parents (44.4%) did not mention any drawbacks of the WSS use in the classroom while a small percentage of the sample (22.2%) held the view that kindergarten teachers procrastinate when they collect, analyze and interpret each child's assessment data.

The presence of parental engagement in the children's assessment at kindergarten is the next research question. Table 8 shows the frequency and the percentage of parents preferring to be engaged in children's assessment at kindergarten.

**Table 8.** Parental engagement in children's assessment

| | Frequency | Percent |
|---|---|---|
| Parents want to get engaged in children's assessment | 10 | 55.6 |
| Parents do not want to get engaged in children's assessment | 8 | 44.4 |
| Total | 18 | 100.0 |

The penultimate research question was designed to explore the views of parents about the reasons why parents should be engaged in the child's assessment or why they should not be engaged. 16.7% of the sample stated that parental engagement in assessment procedures help them be informed about their child's learning development or weaknesses. Also, their engagement urges them to collaborate with kindergarten teachers to solve any problems. Moreover, a small

percentage (5.6%) claims that parents could provide a comprehensive view of their children through home observations, and thus could contribute to the whole process of evaluation. However, 16.7% of the sample is of the opinion that children's assessment implemented by kindergarten teachers is adequately comprehensive and carefully organized. Therefore, parents need not be engaged. Besides, in a few cases (11.2 %), parents do not consider their engagement in children's assessment necessary because either they lack adequate knowledge to evaluate their children or they cannot judge their children objectively.

Finally, regarding parents' views about assessment procedures in kindergarten, the 83.3% of the sample didn't make any statements.

### 3.1. Correlations Between Subscales

The correlation between children's assessment, portfolio assessment, children's developmental checklists and kindergarten teachers' summary reports was checked by Spearman Rank Correlation (rho). The analysis findings are summarized in Table 9.

**Table 9.** Correlations between subscales

| | The role of the portfolio as an evaluation tool in the kindergarten | The role of children's developmental checklists and kindergarten teacher's summary reports |
|---|---|---|
| The role of the implementation of the child's evaluation in preschool education | 0.663 (p = 0.003) | 0.763 (p < 0.001) |
| The role of the portfolio as an evaluation tool in the kindergarten | – | 0.737 (p < 0.001) |

As Table 9 shows, it seems that the role of the implementation of the child's evaluation in preschool education is positively correlated with both the role of the portfolio (Spearman's r = 0.663; p = 0.003) and the role of children's developmental checklists and kindergarten teacher's summary reports (Spearman's r = 0.763; p <0.001). Also, the role of the portfolio, as an evaluation tool in the kindergarten, is positively correlated with the role of children's developmental checklists and kindergarten teacher's summary reports (Spearman's r = 0.737; p <0.001).

### 3.2. Correlations Between Parental Perceptions and their Demographic Factors

The impact of gender, employment status, educational level and household composition on parents' reactions to children's authentic assessment is presented in Tables 10, 11, 12 and 13.

**Table 10.** The impact of gender on parents' reactions

| | | N | Mean | SD | p-value |
|---|---|---|---|---|---|
| The role of the implementation of the child's evaluation in early childhood education | Male | 5 | 33.60 | 1.673 | 0.999 |
| | Female | 13 | 33.08 | 3.252 | |
| The role of the portfolio as an evaluation tool in the kindergarten | Male | 5 | 28.00 | 2.550 | 0.878 |
| | Female | 13 | 27.77 | 2.421 | |
| The role of children's developmental checklists and kindergarten teacher's summary reports | Male | 5 | 33.20 | 2.490 | 0.683 |
| | Female | 13 | 31.54 | 4.701 | |

**Table 11.** The impact of employment status on parents' reactions

| | | N | Mean | SD | p-value |
|---|---|---|---|---|---|
| The role of the implementation of the child's evaluation in early childhood education | Civil servant | 5 | 31.40 | 4.980 | 0.608 |
| | Private employee | 8 | 33.88 | 1.356 | |
| | Self-employed | 4 | 34.25 | 0.957 | |
| The role of the portfolio as an evaluation tool in the kindergarten | Civil servant | 5 | 27.00 | 2.450 | 0.415 |
| | Private employee | 8 | 28.00 | 2.564 | |
| | Self-employed | 4 | 28.75 | 2.500 | |
| The role of children's developmental checklists and kindergarten teacher's summary reports | Civil servant | 5 | 30.60 | 6.427 | 0.927 |
| | Private employee | 8 | 32.38 | 3.292 | |
| | Self-employed | 4 | 32.75 | 3.862 | |

**Table 12.** The impact of educational level on parents' reactions

| | | N | Mean | SD | p-value |
|---|---|---|---|---|---|
| The role of the implementation of the child's evaluation in early childhood education | Post-Compulsory Secondary Education | 12 | 34.00 | 1.207 | 0.317 |
| | University education/Technical Educational Institute | 5 | 31.00 | 4.690 | |
| The role of the portfolio as an evaluation tool in the kindergarten | Post-Compulsory Secondary Education | 12 | 28.42 | 2.275 | 0.382 |
| | University education/Technical Educational Institute | 5 | 26.00 | 1.871 | |
| The role of children's developmental checklists and kindergarten teacher's summary reports | Post-Compulsory Secondary Education | 12 | 33.17 | 2.480 | 0.322 |
| | University education/Technical Educational Institute | 5 | 28.60 | 6.107 | |

**Table 13.** The impact of household composition on parents' reactions

| | Family members | N | Mean | SD | p-value |
|---|---|---|---|---|---|
| The role of the implementation of the child's evaluation in early childhood education | Two people | 2 | 34.50 | 0.707 | 0.628 |
| | Three people | 2 | 33.50 | 0.707 | |
| | Four people | 9 | 33.78 | 1.716 | |
| | Five people | 4 | 31.25 | 5.560 | |
| The role of the portfolio as an evaluation tool in the kindergarten | Two people | 2 | 30.00 | 0.000 | 0.325 |
| | Three people | 2 | 29.50 | 0.707 | |
| | Four people | 9 | 27.22 | 2.729 | |
| | Five people | 4 | 28.00 | 1.828 | |
| The role of children's developmental checklists and kindergarten teacher's summary reports | Two people | 2 | 33.00 | 2.828 | 0.875 |
| | Three people | 2 | 34.00 | 0.000 | |
| | Four people | 9 | 32.56 | 3.468 | |
| | Five people | 4 | 30.50 | 7.048 | |

According to Tables 10, 11, 12 and 13, there is no correlation between the demographic characteristics of the respondents (gender, employment status, educational level and household composition) and the parental reactions to children's assessment in kindergarten.

## 4. DISCUSSION OF FINDINGS

The aim of the present study was to explore the parents' perceptions upon: (a) the significance of child's assessment in preschool education, and (b) the impact of the implementation of the Work Sampling System as an assessment tool on the kindergarten. The majority of the findings of the present study reflect a great parental admission of the significance of children's assessment in preschool education through the use of alternative forms as well as positive attitudes towards the Work Sampling System.

The results of the study revealed that parents acknowledge the function of assessment in kindergarten as an important tool for preschool teachers in order to evaluate the performance and progress of young children. Moreover, most parents find it important to communicate with kindergarten teachers and get feedback regarding their children's progress and learning development. Based on parents' answers, it is obvious that parents identify the major role of authentic assessment in kindergarten and they think it is positive to implement authentic assessment practices in the kindergarten. Similar findings are presented by Rutland and Hall (2013) and Ozturk (2013). Besides, our finding confirms an existing gap in parental views concerning young children's assessment as almost the whole research evidence focuses mostly on the educators' perspectives of the children's assessment, thus setting aside the parental involvement.

Considering the parent's views about the contribution of Work Sampling System portfolio assessment, approximately two thirds of the parents agree that: (a) it helps children to participate energetically and to a great extent in their learning process and reflect on how they acquired knowledge, and (b) it assists teachers, parents and children become aware of their potential or weaknesses in the context of kindergarten setting. Similar findings are also presented and confirmed by the research of Meisels et al. (2010), as parents' ratings indicated the portfolio as an important assessment tool with benefits for everyone who is involved in the evaluation process. Besides, many research studies agree with our finding regarding the meaningful role of using portfolios as an alternative method with preschool children for various pedagogical purposes (Engel, 1993; Gilkerson & Hanson, 2000; Peters, Hartley, Rogers, Smith, & Carr, 2009; Rekalidou, Zantali, & Sofianidou, 2010; Chen & Cheng, 2011; Alacam & Olgan, 2015).

Also, most of the respondents pointed out the importance of children's developmental checklists and kindergarten teachers' summary reports. More specifically, parents concede that both assessment tools that is children's developmental checklists and teachers' summary reports, helped them comprehend the way children think and develop in the kindergarten context. As findings show, nearly 66.7% of parents stated that the WSS tools (except portfolio) are helpful in many ways, considering that: (a) they provide valuable feedback pertaining to the learning process, the level of knowledge, skills or attitudes, potential and possible weaknesses of each child individually in every term and in accordance with the principles and objectives of the kindergarten curriculum, (b) they enlighten possible behavioral problems or learning difficulties of each child, and (c) they give more straightforward information about the kindergarten daily program in general and its cognitive learning areas in particular.

In conclusion, the impact of the implementation of children's developmental checklists and kindergarten teachers' summary reports seems very clear in parents' views. These results are consistent with the findings of relevant surveys that were conducted in the relevant literature (Diffily, 1994; Hannon, 1997; Meisels et al., 2010). It is worth mentioning that the parents paid plenty of attention to those two assessment tools as most of them commented positively on the detailed information they gathered. Parents thought that the new assessment system was especially important to them when it was presented thoroughly in the first term and acknowledged its value, considering it as an appropriate and a valid assessment tool for their children. In the Greek kindergarten, parents are used to informal briefing by kindergarten teachers or parent-teacher group meetings overviewing children's portfolios.

With regard to the advantages of the WSS implementation in the kindergarten, there is a variety of positive opinions in parents' written responses and remarks. The parents observed many benefits in the use of WSS in the kindergarten. The integrated recording of child's progress and development, according to Greek kindergarten curriculum, is recorded as the most common positive advantage. Many individual responses from parents consider the WSS essential in the kindergarten for the following reasons: (a) it involves children in a continuous procedure of development and enhancement of learning process, being at the same time a pedagogical documentation of their learning experiences, (b) it provides a holistic understanding of children's potential, weaknesses and knowledge level and is a diagnostic means of possible learning-behavioral problems, (c) it gives plenty sources of information to parents, and (d) it supports children's self-assessment and encourages them to highlight their needs, interests and efforts in kindergarten everyday activities.

The majority of these positive opinions of parents are justified because children made progress that was noticed through the school year by them as kindergarten teachers used portfolio assessment in the specific kindergarten in the last school year as an alternative method of evaluation. The same findings are presented by the study of Meisels et al. (2010), in which a large percentage of parents (80%) gave high ratings to the use of WSS as well, confirming the benefits for their children.

In order to fully explore parents' views on the drawbacks of the WSS use in kindergarten, responses from four parents indicated that it is a time-consuming process for the kindergarten teacher to collect, analyze and interpret the related data concerning the assessment of each child. Parents strongly realize that collecting the necessary amount of evidence for early learning of children's progress takes a lot of time and is a difficult task for many kindergarten teachers. This result is supported by several studies which identified children's assessment as a complex issue for teachers because they have to provide a valuable profile and document the progress of students investing a lot of time in this significant pedagogical procedure (Appl, 2000; Epstein et al., 2004).

The results also showed that WSS cannot be implemented in kindergarten as an autonomous and an independent assessment tool unless kindergarten teachers organize briefings with parents simultaneously. The implementation of the WSS helps families understand: (a) what assessment is, (b) what the goal of child's assessment is, (c) what kind of alternative assessment methods are used in kindergarten, and (d) what the assessment information means to their child's learning progress and development (Brink, 2002). It is necessary that the portfolio, the developmental checklists and the kindergarten teachers' summary reports - the three main interrelated elements - to function as a whole. Otherwise, the WSS assessment tool cannot be realized to its full extent (Dichtelmiller et al., 2001). Parents need not only an extensive understanding of children's

learning development but also comprehensive knowledge of assessment tools which are appropriate for their children. Thus, this finding confirms other research findings that communication and briefing among educators and parents can be beneficial (Culbertson & Jalongo, 1999; Billman et al., 2005; Murray et al., 2008).

The concept of parental engagement in child's assessment and its role in the overall assessment approach was indicated positively by the respondents. Analyzing the term "parental engagement" in the present study and according to the ratings of the sample, it includes: (a) comprehension of children's learning development, (b) communication with kindergarten teachers, (c) gathering home-based information about children's progress, and (d) collaboration between teachers-parents, which can sort out problems. Similar findings are also presented by Atkinson (2003), Grisham-Brown et al. (2006), and Birbili and Tzioga (2014).

Not surprisingly, parents also describe children's assessment implemented by kindergarten teachers as adequately comprehensive and carefully organized. Therefore, it is not necessary for them to get involved. Parents seemed to be comfortable and satisfied to a great extent with teacher's judgment on child's learning as they became more aware of the children's skills and abilities through the assessment procedure. The same findings are presented in research studies conducted by Meisels et al. (2001) and Shumow (2001). In contrast, slightly less than a third of parents (11.2%) stated that they do not consider their involvement in children's assessment substantial because either they lack adequate knowledge to assess children or they cannot be objective judging their children. In this way, parents affirmed that they have confidence in the kindergarten teacher's role as an assessor. Ultimately, the parental reactions to children's assessment do not vary due to demographics as the findings did not show any positive correlation between them.

## 5. CONCLUSION

Fostering the use of authentic assessment in preschool education is a demanding and at the same time an essential process beneficial for everyone who is involved in this meaningful procedure. What is clear is that all three involved parts namely the child, the teachers and the parents should interact as a useful model and as an integral part of the educational process in early childhood education. This research paper pointed out the views of parents regarding the significance of the child's assessment and the impact of the implementation of the Work Sampling System in a Greek kindergarten.

In this research, the parents considered the implementation of child's assessment useful for everyone who is involved in children's learning and development. The parents seem supportive and satisfied with the forms and functions of the assessment procedure, as they value the children's outcomes throughout the school year. Authentic assessment in kindergarten is equally important as the parents get useful information about children's performance and progress.

The clear message is that assessment in preschool education generally and the implementation of authentic assessment tools more specifically, such as WSS, may well be substantial. The greatest value in authentic assessment lies in children, teachers and parents making use of partnerships to enhance the educational process. Engaging in this type of assessment environment, children, teachers and parents collaborate in an ongoing process that will lead to a greater student learning and personal development.

### 5.1. Limitations and Future Directions

Although this research has reached its aims and yielded some findings, there were some unavoidable limitations. The main limitation of this study is the use of a small number of participants as it does not allow the generalization of the research results and findings. The small population size of the kindergarten and the parental availability, as the whole of them are employees, directed the researchers to implement quantitative research methodology. The implementation of qualitative methods, such as semi-structured group and individual interviews with parents, or a mixed methods research could provide an in-depth analysis and invaluable information of the parental perceptions upon the children's assessment in preschool education. Nevertheless, the present study aims to point out the significance of young children's evaluation in preschool education in general and the parental views in this procedure in particular. Future research and further studies are needed to understand the possible existing gap between parents' beliefs and viewpoints regarding children's authentic assessment in kindergartens. A larger sample would allow for more analyses to determine parents' ratings concerning this particular issue. The views of parents should serve as a starting point for new changes and innovations in assessment of young children in preschool education.

## 6. REFERENCES


Alacam, N., & Olgan, R. (2015). Portfolio assessment: does it really give the benefits that it purports to offer? Views of early childhood and first-grade teachers. *Early Child Development and Care*, Vol. 186, No. 9, 1505-1519.

Appl, D. (2000). Clarifying the preschool assessment process: traditional practices and alternative approaches. *Early Childhood Education Journal*, 27(4), 219-225.

Atkinson, P. (2003). Assessment 5-14: What do pupils and parents think? *Spotlight*, No. 87, 1-4. Edinburgh, UK: The SCRE Centre, University of Glasgow.

Bagnato, S. (2007). *Authentic assessment for Early Childhood Intervention. Best practices.* New York, USA: The Guilford Press.

Bagnato, S., McLean, M., Macy, M., & Neisworth, J. (2011). Identifying instructional targets for Early Childhood via authentic assessment. Alignment of professional standards and practice-based evidence. *Journal of Early Intervention*, Vol. 33, No. 4, 243-253.

Baum, A., & McMurray-Schwarz, P. (2004). Preservice Teachers' Beliefs about Family Involvement: Implications for Teacher Education. *Early Childhood Educational Journal*, Vol. 32, No. 1, 57-61.

Beaton, D., Bombardier, C., Guillemin, F., & Feraz, M. (2000). Guidelines for the Process of Cross-Cultural Adaption of Self-Report Measures. *Spine*, Vol. 25, No. 24, 3186-3191.

Becher, R. (1984). *Parent involvement: A review of Research and Principles of Successful Practice*. (ERIC Document Reproduction Service No. ED247032).

Bergen, D. (1993). Teaching strategies: authentic performance assessments. *Childhood Education*, Vol. 70, Issue 2, 99-102.

Billman, N., Geddes, C., & Hedges, H. (2005). Teacher-Parent Partnerships: Sharing understandings and making changes. *Australian Journal of Early Childhood*, Vol. 30, No. 1, 44-48.

Birbili, M., & Tzioga, K. (2014). Involving parents in children's assessment: lessons from the Greek context. *Early Years*, Vol. 34, Issue 2, 161-174.

Brassard, M., & Boehm, A. (2007). *Preschool Assessment. Principles and Practices*. New York, USA: The Guilford Press.

Brink, M. (2002). Involving parents in Early Childhood Assessment: Perspectives from an Early Intervention Instructor. *Early Childhood Education Journal*, Vol. 29, No. 4, 251-257.

Brodie, K. (2013). *Observation, assessment and planning in the early years.* Berkshire, UK: Open University Press.

Brookhart, S. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record*, Vol. 106, No. 3, 429-458.

Bryman, A. (2016). *Social Research Methods*. Oxford: University Press.

Carr, M. (2001). *Assessment in early childhood settings: learning stories.* London: Paul Chapman.

Chen, S., & Cheng, Y. (2011). Implementing curriculum-based learning portfolio: a case study in Taiwan. *Early Child Development and Care*, Vol. 181, No. 2, 149-164.

Clinton, A., & Guilar, K. (2016). Assessment and Collaboration in Family, Home, and Cultural Contexts. In A. Garro (Ed.), *Early Childhood Assessment in School and Clinical Child Psychology* (pp. 161-182). New York, USA: Springer.

Cohen, M., Manion, L., & Morrison, K. (2008). *Methodology of educational research.* Athens: Metaihmio. [in Greek]

Copple, C., & Bredekamp, S. (Eds.) (2009). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8 (3rd ed.).* Washington, DC: National Association for the Education of Young Children.

Creswell. J. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research.* New York: Pearson.

Culbertson, L., & Jalongo, M. (1999). "But what's wrong with letter grades?" Responding to parents' questions about alternative assessment. *Childhood Education*, 75 (3), 130-135.

Dafermou, Ch., Koulouri, P., & Basagianni, E. (2006). *Kindergarten teacher's guide: educational planning and creative learning environments*. Athens: Organization of Publishing Educational Books. [in Greek]

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education*, Vol. 16, Issues 5-6, 523-545.

Dennis, L., Rueter, J., & Simpson, C. (2013). Authentic assessment: Establishing a clear foundation for instructional practices. *Preventing School Failure: Alternative Education for Children and Youth*, 57(4), 189-195.

Dichtelmiller, M., Jablon, J., Dorfman, A., Marsden, D., & Meisels, S. (2001). *Work Sampling in the Classroom. A Teacher's Manual.* Michigan, USA: Rebus Inc.

Diffily, D. (1994). *What parents think about alternative assessment and narrative reporting.* (ERIC Document Reproduction Service No. ED381230).

Doliopoulou, E., & Gourgiotou, E. (2008). *Evaluation in education with an emphasis on early childhood.* Athens: Gutenberg.

Downs, A., & Strand, P. (2006). Using Assessment to Improve the Effectiveness of Early Childhood Education. *Journal of Child and Family Studies*, Vol. 15, Issue 6, 671-680.

Engel, B. (1993). *Valuing children: Authentic assessment based on observation, reflection and documentation.* (ERIC Document Reproduction Service No. ED368450).

Epstein, A., Schweinhart, L., Debruin, A., & Robin, K. (2004). Preschool assessment: A guide to developing a balanced approach. *Preschool Policy Matters*, 7. Retrieved from: http://nieer.org/resources/policybriefs/7.pdf

Field, A. (2009). *Discovering statistics using SPSS*. London: Sage Publications.

Finello, K. (2011). Collaboration in the assessment and diagnosis of preschoolers: Challenges and opportunities. *Psychology in the Schools*, Vol. 48, Issue 5, 442-453.

Fiore, L. (2012). *Assessment of young children: A collaborative approach*. USA, NY: Routledge.

Frey, B., Schmitt, V., & Allen, J. (2012). Defining authentic classroom assessment. *Practical Assessment, Research & Evaluation*, Vol. 17, No. 2, 1-18. Retrieved from: http://pareonline.net/getvn.asp?v=17&n=2

Galindo, C., & Sheldon, S. (2012). School and home connections and children's kindergarten achievement gains: The mediating role of family involvement. *Early Childhood Research Quarterly*, Vol. 27, Issue 1, 90-103.

Gelfer, J. (1991). Teacher-parent partnerships: Enhancing communications. *Childhood Education*, Volume 67, Issue 3, 164-167.

Gilkserson, D., & Hanson, M. (2000). Family portfolios: Involving Families in Portfolio Documentation. *Early Childhood Educational Journal*, Vol. 27, No. 3, 197-201.

Grisham-Brown, J., Hallam, R., & Brookshire, R. (2006). Using authentic assessment to evidence children's progress toward early learning standards. *Early Childhood Education Journal*, Vol. 34, Issue 1, 45-51.

Gullo, D. (2005). *Understanding assessment and evaluation in early childhood education.* USA, NY: Teachers College Press.

Hall, A., Rutland, J., & Grisham-Brown, J. (2011). Family involvement in the assessment process. In J. Grisham-Brown & K. Pretti-Frontczak (Eds.), *Assessing young children in inclusive settings* (pp. 38-59). Baltimore, MD: Paul H. Brookes Publishing Co.

Hannon, J. (1997). *How will implementing authentic assessment procedures during choice time affect teacher/parent communication?* (ERIC Document Reproduction Service No. ED416955).

Hart, D. (1994). *Authentic assessment: A handbook for educators*. New York: Assessment Bookshelf Series.

Henderson, P. & Karr-Kidwell, P.J. (1998). *Authentic Assessment: An Extensive Literary Review and Recommendations for Administrators.* (ERIC Document Reproduction Service No. ED418140).

Hill, N., & Taylor, L. (2004). Parental School Involvement and Children's Academic Achievement. Pragmatics and issues. *Current Directions in Psychological Science*, Vol. 13, No. 4, 161-164.

Johnson. N. (1993). *Celebrating growth over time: Classroom-based assessment in language arts.* (ERIC Document Reproduction Service No. ED358436).

Kazela, K., & Kakana, D. (2009). Conceptions, views and practices of Greek nursery school teachers about the process of evaluation. *Proceedings of the 7th European Regional Conference on Current issues in preschool education in Europe: Shaping the future* (pp.255-266) [in Greek]. Syros. Retrieved from: http://www.omep.gr/texts/Conference_proceedings_Syros.zip

Lam, T. (2008). A comprehensive approach to assessing children in Early Childhood Education. In P. Grotewell and Y. Burton (Eds.), *Early Childhood Education: Issues and Developments* (pp.223-266). New York, USA: Nova Science Publishers.

Losardo, A., & Notari-Syverson, A. (2011). *Alternative approaches to assessing young children.* Maryland, USA: Paul Brookes Publishing Co.

McDonough, J. and McDonough, S., (1997). *Research Methods for English Language Teachers.* London: Arnold.

Meisels, S.J. (1993). Remaking classroom assessment with the Work Sampling System. *Young Children*, *48*(5), 34-40.

Meisels, S.J. (1997). Using Work Sampling in authentic performance assessments. *Educational Leadership*, 54, 60-65.

Meisels, S.J. (2011). Using Observational Assessment to Evaluate Young Children's Learning: The Technical Quality of the Work Sampling System. Retrieved from: http://www.erikson.edu/wp-content/uploads/AERA-FCD-WSS-summary.pdf

Meisels, S. J., Bickel, D., Nicholson J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teacher judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal*, 38(1), 73-95.

Meisels, S. J., Liaw, F., Forfman, A., & Nelson, R. (1995). The Work Sampling System: Reliability and Validity of a Performance Assessment for Young Children. *Early Childhood Research Quarterly*, 10, 277-296.

Meisels, S., Xue, Y., Bickel, D., Nicholson, J., & Atkins-Burnett, S. (2010). Parental reactions to authentic performance assessment. *Educational Assessment*, 7 (1), 61-85.

Murray, M., Curran, E., & Zellers, D. (2008). Building parent/professional partnerships: an innovative approach for teacher education. *The Teacher Educator*, Vol. 43, Issue 2, 87-108.

NAEYC, 2009. Standards for Early Childhood Professional Preparation. Retrieved from: http://www.naeyc.org/files/naeyc/files/2009%20Professional%20Prep%20stdsRevised%204_12.pdf

Ozturk, M. (2013). Family Partnership in Early Childhood Assessment. *Mediterranean Journal of Social Sciences*, Vol. 4, No. 3, 679-686.

Olmscheid, C. (1999). *Parental involvement: an essential ingredient.* (ERIC Document Reproduction Service No. ED431044).

Orillosa, J., & Magno, C. (2013). Parental involvement in children's assessment in kindergarten. *Educational Measurement and Evaluation Review*, Vol. 4, 47-65.

Osburn, M., Stegman, C., Suitt, L., & Ritter, G. (2004). *Parents' perceptions of standardized testing: its relationship and effect on student achievement.* (ERIC Document Reproduction Service No. EJ848215).

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research and Evaluation*, Vol. 13, No. 4, 1-11. Retrieved from: http://pareonline.net/getvn.asp?v=13&n=4

Peters, M., Seeds, K., Goldstein, A., & Coleman, N. (2008). *Parental Involvement in Children's Education 2007*. Research Report. DCSF RR034.

Peters, S., Hartley, C., Rogers, P., Smith, J., & Carr, M. (2009). Early childhood portfolios as a tool for enhancing learning during the transition to school. *International Journal of Transitions in Childhood*, 3, 4-15.

Rekalidou, G., Zantali, Th., & Sofianidou, M. (2010). Assessment in kindergarten. A pilot project on assessment and self-assessment based on portfolio. *Review of Educational Issues*, (16), 22-38 [in Greek]. Retrieved from: http://www.pi-schools.gr/download/publications/epitheorisi/teyxos16/022-038.pdf

Rutland, J., & Hall, A. (2013). Involving Families in the Assessment Process. *Dialog*, 16 (4), 113.120.

Riley, K., Miller, G., & Sorenson, C. (2016). Early Childhood Authentic and Performance- Based Assessment. In A. Garro (Ed.), *Early Childhood Assessment in School and Clinical Child Psychology* (pp. 95-117). New York, USA: Springer.

Sakellariou, M. (2006). Authentic assessment and daily practice in kindergarten: Educators' and future preschool educators' opinions. *Proceedings of the 5th Panhellenic Conference with theme: Greek Pedagogical and Educational Research.* Thessaloniki: Kyriakidis. [in Greek]

Shepard, L. (1994). The Challenges of Assessing Young Children Appropriately. *The Phi Delta Kappan*, Vol. 76, No. 3, 206-212.

Shepard, L., & Bliem, C. (1995). Parents' thinking about standardized tests and performance assessments. *Educational Researcher*, Vol. 24, No. 8, 25-32.

Shumow, L. (2001). Parents' educational beliefs: Implications for parent participation in school reforms. In S. Redding & L. Thomas (Eds.), *The Community of the school* (pp.205-211). Lincoln, IL: Academic Development Institute. Retrieved from: http://www.adi.org/journal/ss01/chapters/chapter15-shumow.pdf

Snow, C., & Van Hemel, S. (Eds.) (2008). *Early Childhood Assessment: Why, what, and how.* Washington, D.C: The National Academies Press.

Swaffield, S. (2011). Getting to the heart of authentic assessment for learning. *Assessment in Education: Principles, Policy & Practice*, Vol. 18, No. 4, 433-449.

Tuckman, B., & Harper, B. (2012). *Conducting educational research*. Plymouth, UK: Rowman and Littlefield.

Work, W., & Stafford, L. (1987). Parent-teacher communication. *Communication Education*, Vol. 36, Issue 2, 182-187.

Wortham, S. C. (2008). *Assessment in early childhood education*. New Jersey: Pearson.

Yin, R. (1994). *Case Study Research: Design and Methods.* Thousand Oaks, CA: Sage.

## ***Supplementary***

Parents' perceptions towards the role and the function of the child's authentic assessment in preschool education and the impact of the application of alternative forms of assessment such as the Work Sampling System implementation

### Parents' questionnaire

### PART I

**1.1 Please tick [√] the appropriate box:**

**1. Gender:** □ Male □ Female

**2. Age:** □ 19 – 22 □ 23 – 33 □ 34 – 44 □ 45 – 55

**3. Employment status:** □ Civil servant □ Private employee

□ Self-employed □ Unemployed □ Else

**4. Highest educational level:** □ Primary Education □ Compulsory Secondary Education

□ Post-Compulsory Secondary Education

□ University education/Technical Educational Institute

□ Master's degree □ Doctorate degree

**5. Household composition:** □ 1 □ 2 □ 3 □ 4 □ 5 □ > 5

**6. Marital status:** □ Never married □ Married □ Divorced □ Widowed

□ Cohabitation agreement □ Separated

PART II

**2.1 The implementation of children's assessment in preschool education.**

**In this section, indicate the degree to which you agree the statement is important for you. Rate each statement by circling a number between 1 and 5 where the numbers mean the following:**

**1= not at all, 2= a little, 3= enough, 4= a lot and 5= very much**

| **The implementation of children's assessment in preschool education** | NOT AT ALL | A LITTLE | ENOUGH | A LOT | VERY MUCH |
|---|---|---|---|---|---|
| 1. Assessment helps the kindergarten teacher to understand the level of knowledge and skills conquered by children. | **1** | **2** | **3** | **4** | **5** |
| 2. Assessment in kindergarten helps the teacher make instructional design decisions. | **1** | **2** | **3** | **4** | **5** |
| 3. Assessment enables the kindergarten teacher to assess the performance and the progress of young children. | **1** | **2** | **3** | **4** | **5** |
| 4. Assessment in kindergarten assists to record the children's learning development during the school year. | **1** | **2** | **3** | **4** | **5** |
| 5. Assessment in kindergarten facilitates the actual learning of young children. | **1** | **2** | **3** | **4** | **5** |
| 6. Assessment aids the teacher to identify children with learning difficulties or behavioral problems. | **1** | **2** | **3** | **4** | **5** |
| 7. Assessment in kindergarten facilitates briefing of the family. | **1** | **2** | **3** | **4** | **5** |

# PART III

## 3.1 The methods and the techniques of children's authentic assessment using the Work Sampling System.

### 3.1.1 Portfolio assessment

**In this section, indicate the degree to which you agree the statement is true for you. Rate each statement by circling a number between 1 and 5 where the numbers mean the following:**

**1= minimum and 5= maximum**

**Minimum** **1 2 3 4 5** **Maximum**

| **Portfolio assessment** | | | | | |
|---|---|---|---|---|---|
| 1. Helps the children to be involved actively in daily kindergarten learning procedures. | **1** | **2** | **3** | **4** | **5** |
| 2. Helps the children to self-assessment procedure and observe their progress. | **1** | **2** | **3** | **4** | **5** |
| 3. Helps the children to rethink and reflect on how they did their work or how they acquired knowledge. | **1** | **2** | **3** | **4** | **5** |
| 4. Helps the children to develop feelings of autonomy, self-esteem, individual choices and pride. | **1** | **2** | **3** | **4** | **5** |
| 5. Urges the children to express their personal interests, needs and abilities. | **1** | **2** | **3** | **4** | **5** |
| 6. Helps the children, the kindergarten teachers and the parents to assess potential and possible weaknesses. | **1** | **2** | **3** | **4** | **5** |

**3.1.2 Children's developmental checklists and kindergarten teachers' summary reports.**

**In this section, indicate the degree to which you agree the statement is true for you concerning the importance of the children's developmental checklists and kindergarten teachers' summary reports. Rate each statement by circling a number between 1 and 5 where the numbers mean the following:**

**1= minimum and 5= maximum**

**Minimum** **1 2 3 4 5** **Maximum**

| **Children's developmental checklists and kindergarten teachers' summary reports help parents to understand** | | | | | |
|---|---|---|---|---|---|
| 1. The way children think and develop. | **1** | **2** | **3** | **4** | **5** |
| 2. The learning process of each child individually in every period of the school year. | **1** | **2** | **3** | **4** | **5** |
| 3. Children's potential and possible weaknesses. | **1** | **2** | **3** | **4** | **5** |
| 4. Children's progress in accordance with the principles and objectives of the kindergarten curriculum. | **1** | **2** | **3** | **4** | **5** |
| 5. The level of knowledge, skills or attitudes children have acquired. | **1** | **2** | **3** | **4** | **5** |
| 6. The potential behavioral problems or learning difficulties of each child. | **1** | **2** | **3** | **4** | **5** |
| 7. The kindergarten daily program and the cognitive learning areas. | **1** | **2** | **3** | **4** | **5** |

**3.2 An overall evaluation of Work Sampling System as an assessment tool for children.**

**3.2.1 Please mention two benefits and two drawbacks of the WSS application in kindergarten this school year.**

*Benefits*

1……
……..……
……
……..
2……
……...…
……
……

*Drawbacks*

1……
……...…
……..
……
2……
……
……
……

**3.2.2 Do you think that parents should be engaged in the assessment process of their children? If yes, how? If not, why?**

……
……
……
……
……
……
……
……

**3.2.3 If you have additional opinion or remark concerning the process of kindergarten assessment that has not been reported, please mention:**

……
……
……
……
……
……

Thank you very much for your participation and cooperation.

*Int. J. Asst. Tools in Educ., Vol. 4, Issue 2, (2017) pp. 211-222*

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

*"Research Article"*

# Using the 2006 PISA Questionaire to Evaluate the Measure of Educational Resources: A Rasch Measurement Approach

**Ruixue Liu**[*,1], **Letao Sun**[1], **Jing Yuan**[1], **Kelly Bradley**[1]

[1]University of Kentucky, United States

***Abstract***

School educational resources are key when studying school improvement due to their influence on learning outcomes. Because of this, careful attention should be given to the way educational resources are operationalized and measured. Using the 2006 PISA American sample containing 166 schools, this study aims to validate the 13-item PISA School Educational Resource Scale with Rasch analysis. Winsteps software was used in the analysis and results were used to evaluate how well the instrument measured the construct of school educational resource. Findings revealed that the PISA 2006 data gave an overall indication of good fit to the model, despite the instrument not separating respondents well. In regards to the quality of the scale, the majority of items perform consistently with the model. However, for schools above the average educational resource threshold, it appears there is a need for more items to discriminate the situation.

***Article Info***

***Received***
*February 15, 2017*

***Revised***
*May 27, 2017*

***Accepted***
*June 01, 2017*

***Keywords***

School educational resource,
PISA,
Rasch rating scale model,

## 1. INTRODUCTION

According to Hanushek (1997), school educational resource was operationalized as the combination of the real resources of the classroom (e.g. teacher education, teacher experience, and teacher-pupil ratios), financial aggregates of resources (e.g. expenditure per student and teacher salary), and estimates of other resources in school (e.g. specific teacher characteristics, administrative inputs, and facilities). School educational resource plays a critical role in attaining educational objectives and create equal opportunities for students (Savasci & Tomul, 2013). With the Every Student Succeeds Act (ESSA), the federal government has become more deeply involved in seeking to improve student achievement. With the emphasis on the development of

[*]Corresponding Author E-mail: liuruixue2046@hotmail.com

2148-7456 /© 2017 DOI: 10.21449/ijate.319486

student achievement, educational leaders and policymakers should make effective decisions on allocating school educational resource to help school meet student learning objective. To make these decisions, educational leaders and policymakers need reliable evidence of the effects of specific educational resources on student achievement (Sala, 2014).

This study applied the Rasch rating scale model to assess the quality of the School Educational Resource Scale, an instrument used to evaluate school educational resources in Program for International Student Assessment (PISA) 2006. Specifically, the aim of the study is to provide an overall assessment of the psychometric properties of this instrument. Findings may lead to a more accurate measure of school educational resources.

### 1.1. School Effectiveness Research

Studies of school educational resources have been embedded in school effectiveness research (Murnane, 1981; Schneider, 1985; Ma, 2001; Konstantopoulos, 2006; Stanco, 2012).

An effective school has been defined in different ways (Johnson, 2008). For example, Lezotte (2001) claimed that an effective school should provide *"(1) instructional leadership, (2) clear vision and mission, (3) safe and orderly environment, (4) high expectations for student's achievement, (5) continuous assessment of student achievement, (6) opportunity and time on task and (7) positive home-school relations*" (p.4). Some researchers have focused on academic achievement of the students (e.g., MacNeil, Prater, & Busch, 2009; Koth, Bradshaw, & Leaf, 2008), while other researchers concentrated on differences in attitudes and behavior of the students (e.g., Elliot, Cornell, Gregory, & Fan, 2010; Way, Reddy, & Rhodes, 2007).

The following effective school definition was adopted by the Organization for Economic Cooperation and Development ([OECD], 1994) with a global approach: "*An effective school promotes the progress of its students in a broad range of intellectual, social, and emotional outcomes, while considering socio-economic status, family background and prior learning*" (p.1).

School effectiveness studies covered three generations over the past 50 years (Fan, 2013). The first generation of school effectiveness research started about 50 years ago with the publication of Coleman and his colleagues' (1966) research on the quality of schooling in the United States. This study, known as *The Coleman Report*, has been regarded as the first large-scale study of school effectiveness and considered as the major impetus for development of research in this field (Reynolds, Creemers, Stringfield, Teddlie, & Schaffer, 2002). In this study, the results of standardized test of ability and achievement for a total of 645,000 students from more than 4,000 schools were collected and analyzed to explore whether the schools had a measurable impact on student achievement. Coleman et al. concluded that schools have relatively little impact on student achievement compared to the socioeconomic background and started an ongoing debate.

A group of noteworthy school effectiveness studies in the mid-1980s, including the *School Matters* in London (Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988) and *Louisiana School Effectiveness Study* (LSES) (Teddlie & Stringfield, 1993), were considered the second generation of school effectiveness studies. In the study of *School Matters*, Mortimore et al. (1988) aimed to examine the size of school effect, differentiate school effectiveness, and identify factors to develop school effectiveness. Two thousand children, randomly selected from 50 primary schools participated in this study over the course of four years. The LSES was a longitudinal study conducted from 1980 to 1992, utilized both quantitative and qualitative methods (Teddlie et al. 1993) in the United States. This was a longitudinal study from 1980 to 1992 which utilized both

quantitative and qualitative methods to analyze data at the school and classroom levels. Several factors to promote effectiveness of middle school with low SES were discovered and discussed, including the enhancement of educational expectations; principal leadership style; usage of external reward structures; the emphasis on school curriculum; parental involvement; and the experience level of teachers.

In the third decade, the school effectiveness research shifted toward a globalization in the field (Teddlie & Reynolds, 2000). The majority of school effectiveness studies have been conducted in the western countries such as the United States, the United Kingdom, the Netherlands, Australia and Canada. As Teddlie (2004) called attention to and is still the case today, it is necessary to also study under-represented areas of the world to enrich the knowledge base of this field and to make comparisons with the existing research.

### 1.2. School Educational Resources

Many studies have researched the question of whether the level, or amount, of school educational resources influenced student outcomes of learning. Unfortunately, it has proven difficult to determine the relationship between school educational resources and student achievement outcomes (Sala, 2014). According to Hanushek (1997), evidence was not found to support a strong or consistent relation between school educational resources and student achievement. This finding has received considerable attention and acceptance by individuals in the academic, legal, and public policy arenas. Others have challenged this position and results from other studies provide counter evidence. Knoeppel, Verstegen, and Rinehart (2007) found that average school wealth has positive effects on student achievement. Moreover, Jacob and Ludwig (2008) showed that increased funding used in early childhood education, class size reduction, and salary lead to improved student outcomes. Vandiver (2011) indicated that quality and educational adequacy of educational facilities were statistically significantly correlated with student performance.

According to the Organisation for Economic Co-operation and Development (2010), effective school systems require the right combination of qualified personnel, adequate educational resources, facilities, and motivated students ready to learn; in addition, factors including class and school size, the quality of teaching materials, perceived staff shortages, and teacher quality are frequently associated with student performance. Most noticeably, school educational resources are the most important set of mediators through which the socio-economic background of students and schools affects performance.

The mixed findings on the effectiveness of school educational resource on academic achievement may partly due to instruments with an inadequate quality. Thus, it is necessary to develop a more reliable and valid instrument to measure school educational resources. The Rasch model, as a powerful approach to investigate psychometric properties, was conducted in this study. The following section will provide a brief introduction of Rasch model.

### 1.3. Rasch Model

According to Wright and Linacre (1989), the arithmetical property of interval scales is fundamental to any meaningful measurement. Traditional analytical techniques usually anchor on True Score Theory, and the raw data are not interval data. Thus, the data only indicate ordering without any proportional meaning (Yan & Mok, 2012). According to Waugh and Chapman (2005), one cannot make valid inferences from the measures that are initially set up for True Score Theory.

The aforementioned issue can be overcome by analyzing the data via the Rasch model. The Rasch model, introduced by Georg Rasch (1960), can generate a comprehensive picture of the association between observed item responses on a scale and persons' levels on a latent variable. The Rasch model is the simplest of the Item Response Theory (IRT) models, having a single parameter for the person or entity and a single parameter corresponding to each category of an item. An application of the Rasch model is appropriate any time a researcher wishes to use the total score on an assessment or questionnaire to make inferences about an individual's ability or level of a latent trait inherent in that individual (Bond & Fox, 2001).

Since the Rasch model arises from the requirement that comparisons among person and items are invariant across samples, it is appropriate when the total score on a test or questionnaire is used to make inferences. Although Classical Test Theory (CTT) also uses the total score to characterize each person, the total score is used as the relevant statistics without paying enough attention on the anomalies in the items or persons answering them. These anomalies can be explained by the Rasch model which can provide a more informative score. The objective of Rasch measurement is similar with the construction of a ruler, establishing the correct measure (Andrich & Luo, 2003).

The Rasch model is a methodological tool that can be used to analyze data, especially when dealing with latent traits such as attitudes or perceptions. It allows observations of respondents and items to be connected in a way that indicates the occurrence of a certain response as probability rather than certainty and maintains order in that the probability of providing a certain response defines an order of respondents and items. In other words, a person endorsing an extreme statement, or answering a difficult item, should also endorse all less extreme statements, or answer correctly the less difficult items (Wright & Masters, 1982). A rating scale is a set of categories designed to elicit information about a quantitative or a qualitative attribute. In the social sciences, a common example is the use of a Likert scale in which a person selects the number which they consider to reflect the perceived quality of a product (Andrich, 1978). In the current study, the rating scale model was used, as it is appropriate for the analysis of survey data. The formula is:

$$\ln\left(\frac{Pnij}{Pnij}\right) = Bn - Di - Fj \qquad 1$$

In Equation 1, *Pnij* = the probability that person *n* encountering item *i* is observed in category *j*, *Bn* = the "ability" measure of person *n*, *Di* = the "difficulty" measure of item *i*, (the point where the highest and lowest categories of the item are equally probable), *Fj* = the "calibration" measure of category *j* relative to category *j*-1 (Rasch-Andrich threshold located at the point of equal probability of categories j–1 and j); and no constraints are placed on the possible values of *Fj*. Winsteps measurement software was used to perform the Rasch analysis (Linacre, 2009).

## 2. METHOD

### 2.1. Data Source

The primary database used in this research is constructed from the Program for International Student Assessment (PISA) conducted in 2006. According to Organization for Economic Co-operation and Development (OECD) (2001), PISA is the most comprehensive and rigorous international assessment on 15-year-old student performance in reading, science, and mathematics.

Every three years, data is collected on the student, family and institutional factors that is used to analyze differences in performance. PISA examines how well students are prepared to meet the challenges of the future and how well students are prepared for life in a larger context, rather than how well they master particular curricula. In 2006, PISA included information on nearly 400,000 students from 57 countries. The database included student performance in reading, science, and mathematics. In addition, data from the parents and school principals of participating schools were also included.

The data for this study is derived from the United States sample in the 2006 PISA study conducted by OECD. Data were downloaded from the OECD website. SPSS 22.0 program was used to manage and clean the data. The sample contains 166 persons (high school principals). Eleven persons who failed to complete this survey were excluded from the Rasch analysis. Therefore, there were 155 persons measured on the 13 items for this study.

### 2.2. Instrument

The entire set of items used in this scale is derived from the school questionnaire of PISA 2006. The index of school educational resource aims to measure principals' perceptions of potential factors hindering instruction at schools through the 13-item scale (e.g., a lack of qualified science teachers; shortage or inadequacy of science laboratory equipment; shortage or inadequacy of computer software for instruction; Shortage or inadequacy of audio-visual resources). A four point Likert-type scale was used (not at all = 1, very little = 2, to some extent = 3, a lot = 4). As all items were inverted for scaling, higher values on this index indicate more school educational resources. The detailed items can be found in Table 1.

**Table 1.** Items of School Educational Resource Assessment

| Question | Items | Responses |
|---|---|---|
| Is your school's capacity to provide instruction hindered by any of the following? | 1. A lack of qualified science teachers<br>2. A lack of qualified mathematics teachers<br>3. A lack of qualified (test language) teachers<br>4. A lack of teachers of other subjects<br>5. A lack of laboratory technicians<br>6. A lack of other support personnel<br>7. Shortage or inadequacy of science laboratory equipment<br>8. Shortage or inadequacy of instructional materials<br>9. Shortage or inadequacy of computers for instruction<br>10. Lack or inadequacy of Internet connectivity<br>11. Shortage or inadequacy of computer software for instruction<br>12. Shortage or inadequacy of library materials<br>13. Shortage or inadequacy of audio-visual resources | 1 - Not at all<br>2 - Very little<br>3 - To some extent<br>4 - A lot |

## 3. RESULTS

*Dimensionality Analysis:* The Rasch principal components analysis of residuals was carried out to assess the dimensionality of the constructed scale. The eigenvalue of the first contrast was 3.3, indicating it has the strength of about three items (3.3 rounded to 3, out of 13). It is larger than the strength of two items (an eigenvalue of 2), the smallest amount that could be considered a dimension. Meanwhile, the eigenvalue of second contrast is 1.8. Thus the assumption of unidimensionality holds, and is not violated, in this study.

*Reliability and Separation:* Both reliability and separation statistics can be considered at the person and item level. Person reliability is analogous to Cronbach's alpha reliability in True Score Theory while item reliability has no traditional equivalent. Low values for item reliability indicate a narrow range of item measures, or a small sample. Person separation is used to classify people, and item separation is used to verify the item hierarchy (Linacre, 2009). The reliability and separation statistics can be found in Table 2. Person reliability was computed to be 0.76, and item reliability was 0.90. Person separation was 1.76, and item separation was 3.07.

**Table 2.** *Model Fit Statistics*

| | Measure | Infit ZSTD | Outfit ZSTD |
|---|---|---|---|
| Principals (Reliability= .76; Real RMSE=.70) | | | |
| *M* | 42.20 | -.10 | .00 |
| *SD* | 6.60 | 1.30 | 1.30 |
| Items (Reliability= .90; Real RMSE=.13) | | | |
| *M* | 502.80 | -.10 | -.10 |
| *SD* | 27.70 | 1.90 | 1.60 |

*Model Fit Statistics:* ZSTD is a *t*-test of the hypothesis "Do the data fit the model (perfectly)?" They are reported as z-scores. Besides, they show the improbability of the data, if the data actually fits the model. Zero are their expected values. Less than 0 indicates too predictable. More than 0 indicates lack of predictability. Generally, if the ZSTD were within the range of -1.9 to 1.9, the instrument indicates a reasonable predictability (Linacre, 2002). Table 2 showed that both the infit and outfit ZSTD could meet this requirement.

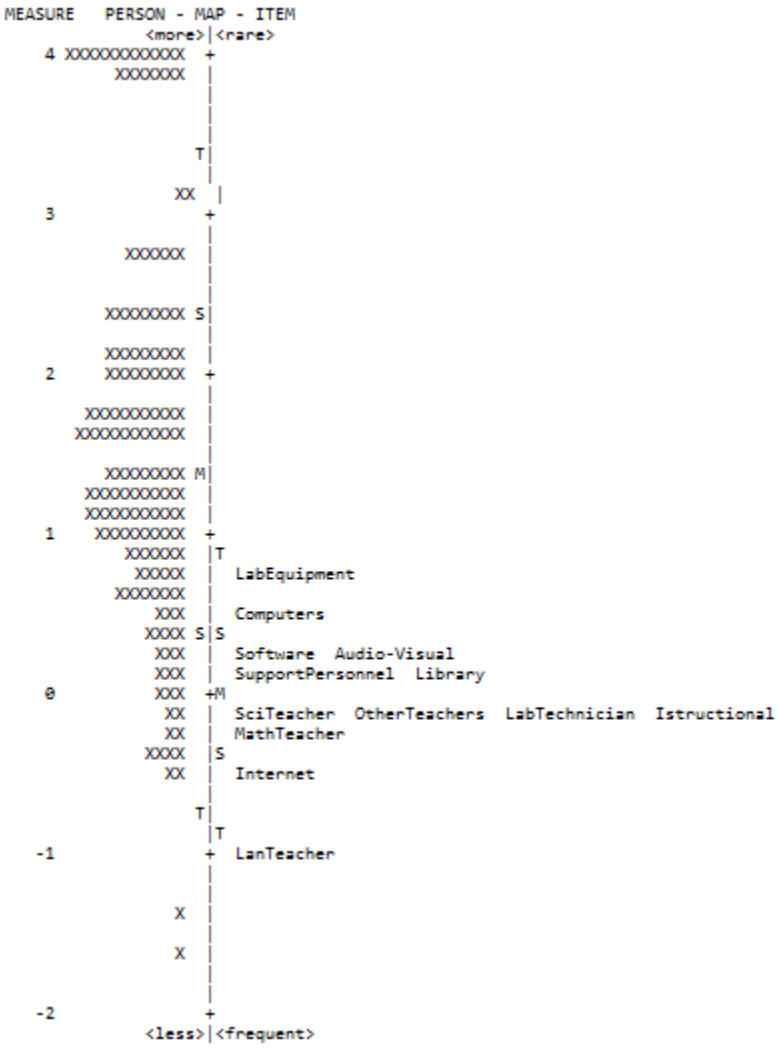
*Item Infit and Outfit:* There are two types of item fit statistics in the Rasch analysis. Item outfit statistics are influenced by unexpected responses to items, for example, when a person of low ability gets a very difficult item correct. Infit statistics are influenced by an unexpected pattern of responses near a person's ability estimate, for example, when a person gets the item near the person's ability estimate incorrect.

Table 3 shows the item misfit statistics, which reveals several misfitting items. For instance, Item 2 (A lack of qualified mathematics teachers) has the maximum infit indices (ZSTD = 3.50), which exceed the upper bound of criteria range of infit ZSTD (3.50 > 2), and Item 11 (Shortage of inadequacy of computer software for instruction) has the minimum infit indices (ZSTD = -3.20), which exceed the lower bound of criteria range of infit ZSTD (-3.2 < -2). In addition, Item 11 also has the minimum outfit indices (ZSTD = -2.60) that exceed the lower bound of criteria range of outfit ZSTD (-2.60 < -2) (see Table 3).

**Table 3.** *Item Statistics*

| Items | Measure | Infit | Outfit |
|---|---|---|---|
| | | ZSTD | ZSTD |
| 2 | .23 | 3.50 | 2.80 |
| 1 | .07 | 2.20 | 1.60 |
| 5 | .12 | 1.90 | 1.70 |
| 3 | .94 | .70 | .90 |
| 6 | -.18 | 1.00 | 1.30 |
| 9 | -.50 | .90 | 1.00 |
| 7 | -.78 | -.40 | -.20 |
| 4 | .17 | -.50 | -.040 |
| 10 | .44 | -1.10 | -1.10 |
| 13 | -.23 | -1.80 | -1.80 |
| 12 | -.14 | -1.90 | -1.70 |
| 8 | .12 | -2.40 | -2.20 |
| 11 | -.29 | -3.20 | -2.60 |

**Figure 1.** *Item-person map for school resource items*



*Item and Person Map:* Figure 1 shows the item-person map, which provides distribution for both item difficulty and person ability estimates on a single line of logit scale to facilitate the graphical representation of the relationships. This map displays the person measure and item measure on the same scale. The ability estimates are shown on the left side and the item difficulty locations are shown on the right. Person ability and item difficulty increase as one moves towards the top of the figure (Linacre, 2009). Overall, this map shows that the majority of person ability distribution falls outside of the range of the item difficulty distribution. Persons' ability scoring around 0 logits are found to be well measured by the items, and all item difficulty estimates are clustered around 0 logits. However, the ability distribution is higher overall than the difficulty distribution, which indicates that persons with higher ability are not accurately, or maybe fully, measured by the items.

## 4. DISCUSSION

The item separation and reliability statistics showed that the person sample is large enough to confirm the item difficulty hierarchy (construct validity) of the instrument. However, low person separation (less than 2) and person reliability (less than 0.8) implied that the instrument may not be sensitive enough to distinguish between high and low performers. Adding more items could be a solution to the issue. Meanwhile, the analysis of misfit reveals some potentially misfitting items on the school educational resource scale, suggesting revision may be needed. The item-person map reveals that persons with higher ability are not accurately measured by the items.

The central focus of school effectiveness research concerns the idea that "*schools matter, that schools do have major effects upon children's development and that, to put it simply, schools do make a difference*" (Reynolds & Creemers, 1990, p. l). Moreover, as mentioned earlier, many studies have examined the question of whether the level, or amount, of school educational resources influences the level, or outcomes, of student learning. Some studies indicate that school educational resources of do not have an effect on academic achievement of students (Hanushek, 1997; Hanushek & Luque, 2003). On the other hand, some studies say the exact opposite (Card & Krueger, 1996; Greenwald et al., 1996). This debate leads to researchers seeking instruments to measure school educational resources. With so many instruments, some of them may not be high quality measures, illustrating poor quality in terms of the reliability and validity. Instruments with low reliability may produce different results under comparable, consistent conditions. Validity can help determine what types of assessments to use and make sure whether a method can truly measure the idea or construct in question. Because of this, careful attention should be given to the way educational resource is operationalized and measured and developing a more reliable and valid instrument to measure school educational resources may be the most important part of conducting a high quality research study in this area.

Above all, using a powerful technique to evaluate the psychometric properties of an instrument is important. The current study evaluated how well the instrument measured the construct of school educational resources by analyzing the constructed scale. A good Likert-type scale is grounded in sufficient items with a varying degrees of difficulty to evaluate a range of abilities held by the persons. Utilizing the Rasch model to analyze survey research data will result in more sound measures and more meaningful results (Bond et al., 2001). For example, the Rasch model produces estimates of the latent trait displayed by each subject ("person measure") and the trait to respond in a certain way to each item ("item measure"). The Rasch model also provides item fit statistics that indicate whether the individual item is contributing to the measurement of

the latent trait (Bond et al., 2001). Furthermore, the Rasch model software (e.g., Winsteps) can provide indices and visual displays that help examine whether items and persons spread sufficiently along the continuum of the measure (Linacre, 2009). This enables survey researchers to visualize if and where additional items are necessary to cover the entire dimension of the construct. Above all, researchers and practitioners in testing and measurement should be aware of the advantages of using Rasch analysis.

## 5. CONCLUSIONS

In the current study, the Rasch analysis' results provide a more detailed and comprehensive display of how school principals perceive potential factors hindering instruction at their schools. These results could be disseminated to provide PISA administrators with useful information to make more informed decisions regarding survey administration methods and the interpretation and comparability of the impending results. By using the same framework, the Rasch analysis can be used to examine other school context and climate variables (e.g., teacher effectiveness, classroom practice, and principal leadership) in the school effectiveness research, large-scale assessment, and international comparative studies.

The results of this study, which employed the Rasch measurement model to analyze the PISA 2006 data, give an overall indication of good fit to the model. There were two major weaknesses of the instrument brought to light through this analysis. On the one hand, the item-person map and the statistics of person separation and reliability indicate that there are not enough items to discriminate the situation of school educational resources for schools that are above the average. Even so, this might not matter, as those above average might have reached a successful plateau. On the other hand, some misfitting items were discovered by the analysis of misfit, and they are suggested to be revised in the future research.

The alignment between accountability policies and school finance policies to better serve student learning goals has been emphasized by educational researchers (Superfine, 2009). Findings of this study can contribute to the future research on the effects of school educational resources on student academic achievement. To this end, educational policymakers will have reliable evidence of school educational resources to inform resource allocation practices to meet the demands of educational adequacy.

## 6. REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 357-74.

Andrich, D. & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response Categories using principal components. *Journal of Applied Measurement, 4*(3), 205-221.

Ammermueller, A., Heijke, H., Woessmann, L. (2005). Schooling quality in Eastern Europe: Educational production during transition. *Economical Educational Review, 24* (5), 579 599.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Card, D., & Krueger, A. (1996). School resources and student outcomes: An overview of the literature and new evidence from North and South Carolina. *Journal of Economic Perspectives*. doi: jep.10.4.31

Coleman, J. S., Hoffer, C., & York, R. (1966). *The equality of educational opportunity study*. Washington, DC: United States Department of Health, Education, and Welfare.

Dodson, C. K. (2005). *The Relationship between School Effectiveness and Teachers' Job Satisfaction in North Mississippi Schools.* Unpublished Doctoral Dissertation, Mississippi University, Oxford.

Eliot, M., Cornell, D., Gregory, A., & Fan, X. (2010). Supportive school climate and student willingness to seek help for bullying and threats of violence. *Journal of School Psychology, 48*, 533-553. doi:10.1016/j.jsp.2010.07.001

Fan, M. (2013). *Stability of academic performance across science subjects among Chinese students* (Unpublished master's theses). University of Kentucky, Lexington, KY.

Hanushek, E. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141-164. doi:10.2307/1164207

Hanushek, E. A., & Luque, J. A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review, 22*, 481-502. doi: 10.1016/S0272-7757(03)00038-4

Knoeppel, R. C., Verstegen, D. A., & Rinehart, J. S. (2007). What is the relationship between resources and student achievement: A canonical analysis. *Journal of Education Finance, 33*(2), 183-202.

Jacob, B. & Ludwig, J. (2008). *Improving Educational Outcomes for Poor Children*. Cambridge, MA: National Bureau of Economic Research.

Johnson, A. D. (2008*). The relationships among middle school student and staff perceptions of school effectiveness and student achievement*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Koth, C., Bradshaw, C., & Leaf, P. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology, 100*(1), 96-104. doi: 0022-0663.100.1.96

Lezotte, L. W. (2001). *Revolutionary and Evolutionary: The Effective Schools Movement. Retrieved from http://www.edutopia.org/pdfs/edutopia.org-closing-achievement-gap lezotte-article.pdf.*

Linacre, J. M. (2002).What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2009). A user's guide to Winsteps, Ministep, Rasch-model computer programs: Program manual 3.72.3. Retrieved from http://www.winsteps.com/a/winsteps-manual.pdf.

Ma, X. (2001). Stability of school academic performance across subject areas. *Journal of Educational Measurement, 38*(1), 1-18.

MacNeil, A., Prater, D., & Busch, S. (2009). The effects of school culture and climate on student achievement. *International Journal of Leadership in Education, 12*(1), 73–84. doi: 10.1080/13603120701576241

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters*. Berkeley, CA: The University of California Press.

Murnane, R. (1981). Interpreting the evidence on school effectiveness. *Teachers College Record, 83*(1), 19-35.

Organization for Economic Cooperation and Development. (1994). *Making education count: Developing and using international indicators*. Paris: Author.

Organization for Economic Cooperation and Development (2010). *PISA 2009 Results: What Makes a School Successful? – Resources, Policies and Practices (Volume IV).* Paris: Organization for Economic Cooperation and Development.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Reid, K., Hopkins, D., & Holly, P. (1987). *Towards the effective school.* Oxford: Blackwell.

Reynolds, D & Creemers, B. (1990). School effectiveness and school improvement: A mission statement, *School Effectiveness & School Improvement, 1*(1): 1-3. doi: 10.1080/0924345900010101

Reynolds, D., Creemers, B., Stringfield, S., Teddlie, C., & Schaffer, G. (2002). *World class school: International perspectives on school effectiveness.* London: Routledge Farmer.

Savasci, H. & Tomul, E. (2013). The relationship between educational resources of school and academic achievement. *International Education Studies, 6*(4), 114-123. doi:10.5539/ies.v6n4p114

Sala, M. (2014). *Examining the effects of school-level variables on elementary school students' academic achievement: The use of structural equation modeling* (Unpublished doctoral dissertation). Clemson University, Clemson, SC.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions, 10*, 516-517.

Schneider, B. (1985). Further evidence of school effects. *Journal of Educational Research, 78*(6), 351-356.

Stanco, G. (2012). *Using TIMSS 2007 data to examine STEM school effectiveness factors in an international context.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Superfine, B. M. (2009). Deciding who decides questions at the intersection of school finance reform litigation and standards-based accountability policies. *Educational Policy 23*(3), 480-514. doi: 10.1177/0895904808314712

Teddlie, C., & Stringfield, S. (1993). *Schools make a difference: lessons learned from a 10-year study of school effects.* New York: Teachers College Press.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press.

Teddlie, C. (2004). Getting schools working in South Africa: A valuable addition to the SESI field. *School Effectiveness and School Improvement*, *15*(2), 227-240.

Vandiver, B. (2011). *The impact of school facilities on the learning environment* (Doctoral dissertation). Retrieved from ProQuest Dissertations Publishing (UMI No. 3439537)

Waugh, R. & Chapman, E., (2005). An analysis of dimensionality using factor analysis (true score theory) and Rasch measurement: what is the difference? Which method is better? *Journal of Applied Measurement, 6*(1), 80-99.

Way, N., Reddy, R., Rhodes, J. (2007). Students' perceptions of school climate during the middle school years: Associations with trajectories of psychological and behavioral adjustment. *American Journal of Community Psychology, 40*, 194–213.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal: Measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*(12), 857-860.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Yan, Z. & Mok, M. (2012). Validating the coping scale for Chinese athletes using multidimensional Rasch analysis. *Psychology of Sport and Exercise, 13*, 271-279. doi:10.1016/j.psychsport.2011.11.013.