APPLICATION OF RASCH MEASUREMENT THEORY TO THE TORRANCE TESTS OF CREATIVE THINKING, FIGURAL FORM A

by

SUREYYA YORUK

(Under the Direction of Bonnie Cramond)

ABSTRACT

Five overarching assumptions are currently made about the items in the Torrance Tests of Creative Thinking, Figural Form A (TTCT-figural Form A; Torrance, 1966). The items are assumed to have a good fit, to be equally difficult, to be composed of equally distributed response options, to behave the same for each gender, and to involve the same amount of standard error of measurement which is at the same level along the latent trait continuum. In the present study, Rasch measurement theory was utilized to examine the items in the TTCT-figural Form A for each of these assumptions regarding all the variables measured in the test, as well as item-level elaboration (i.e., elaboration-I, a variable fabricated for the present study). Data were collected from 193 second grade students in Turkey. The dichotomous Rasch model and the rating scale model were used for analyses. It was found that all the items had a good fit regarding 14 variables. Misfit was detected for three items in total regarding three variables. The items were not estimated to be at the same difficulty level regarding any of the variables except elaboration-I and humor. The response options of the items (or activities) were distributed evenly regarding resistance to premature closure but not elaboration, elaboration-I, and abstractness of titles. The items were estimated to behave the same for each gender regarding the majority of the

variables. Differential item functioning was detected for eight variables. Finally, it was estimated that all the items involved the same amount of standard error of measurement and that the amount of standard error of measurement involved in an item increased towards the far ends of the latent trait continuum. Overall, the findings suggested that the items in the TTCT-figural Form A possessed sufficient quality for providing appropriate person measures for both genders regarding the majority of the variables. The item reliability indexes for 14 variables indicated a high possibility that invariant item calibration was attained regarding those 14 variables. Invariant measurement of the trait levels of the students, on the other hand, was not achieved regarding any of the variables.

INDEX WORDS: Torrance figural tests, Rasch measurement theory, item fit, item difficulty, item scaling, differential item functioning, item information, standard error of measurement, invariant measurement

APPLICATION OF RASCH MEASUREMENT THEORY TO THE TORRANCE TESTS OF CREATIVE THINKING, FIGURAL FORM A

by

SUREYYA YORUK

B.A., Istanbul University, Turkey, 2006

M.Ed., The University of Georgia, U.S.A., 2013

A Dissertation Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

APPLICATION OF RASCH MEASUREMENT THEORY TO THE TORRANCE TESTS OF CREATIVE THINKING, FIGURAL FORM A

by

SUREYYA YORUK

Major Professor: Bonnie Cramond

Committee: George Engelhard, Jr.
Mark A. Runco
Selcuk Acar

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

DEDICATION

To my mother who carried me for nine months and afterwards took care of me for several years, and to my father who constantly supported us and provided us a home.

ACKNOWLEDGEMENTS

First and foremost, I thank God—praise be to Allah—for this opportunity and for giving me necessary power, health, and patience to finish my Ph.D.

Without the scholarship I was given by the Turkish government, it would have been impossible for me to pursue a Ph.D. degree in the U.S. and to write this dissertation. I would like to express my deepest appreciation to those who initiated this scholarship program and to those who made it available to numerous students in Turkey.

I would like to thank my parents and my siblings who supported me and gave me courage along this journey. I also would like to thank my friends, both in Turkey and in the U.S., who always gave me extra energy to keep up with my studies. I especially thank Burak Turkman for scoring some of the tests to check the interrater reliability.

I would like to thank my advisor, Bonnie Cramond, and the members of my dissertation committee, Mark A. Runco, George Engelhard, and Selcuk Acar, for the guidance and inspiration they provided.

Finally, I would like to thank everyone who came into my life at some point in the past 32 years. Whether they were in my life for a short conversation or for a long-lasting friendship, they all impacted me in a certain way and made me who I am today.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Assessment of the potential for creativity has been a primary research interest for years (e.g., Mednick, 1968; Runco, 2001; Torrance, 1966, 1984; Wilson, Guilford, & Christensen, 1953). Several types of instruments, such as scales, questionnaires, and tests, were developed for this particular purpose. Of these instruments, the Torrance Tests of Creative Thinking, Figural Forms A and B (TTCT-figural Forms A and B; Torrance, 1966, 1984) are among the most prominent. Both in the U.S. and outside of the U.S., the TTCT-figural are commonly used in academia for creativity research (Aslan, 2001; Aslan & Puccio, 2006; Krumm, Lemos, & Filippetti, 2014; Palaniappan & Torrance, 2001) and in K-12 education for the identification of creatively gifted students (Hunsaker, Abeel, & Callahan, 1991; Kaufman, Plucker, & Russell, 2012).

The TTCT-figural were first published in their current form in 1966 (Torrance, 1966). Since then, a considerable amount of research has been conducted on the tests. Effective test administration (e.g., Hattie, 1980), reliability of the test scores (e.g., Torrance, 1998), validity of the test scores (e.g., Torrance, 1981), fairness of the test scores (e.g., Matud & Grande, 2007), and standardization of the tests in different cultures (e.g., Aslan & Puccio, 2006) were the objectives of various studies. Despite this voluminous research on the TTCT-figural, no study has focused on the properties of the test items (e.g., item fit and item difficulty) and has investigated the quality of the tests conducting item-level analyses.

Because previous studies on the TTCT-figural used statistical methods based on total test score analyses and did not examine the individual test items, there is currently no statistical evidence that identifies the properties and quality levels of the items in the TTCT-figural. The purpose of the current study was to conduct Rasch measurement theory analyses to examine the individual items in the TTCT-figural Form A in terms of item fit, item difficulty, average item difficulty, item scaling, differential item functioning, item information, and invariant (i.e., sample-independent) calibration of the items as well as invariant (i.e., item-independent) measurement of the creative thinking skill levels of examinees.

Definition of Terms

Several terms are used repeatedly throughout the paper. These terms will be defined below.

Activity: Activity refers to each subtest in the TTCT-Figural.

Item: Item refers to each stimulus in the TTCT-figural. Each stimulus is either an incomplete figure or a geometrical shape (i.e., parallel line or circle).

Variable: Creative thinking skill measured in the TTCT-figural. Currently, 18 different variables are measured in the TTCT-figural. For the present study, a 19^{th} variable (i.e., elaboration-I) was fabricated. Elaboration-I is item-level elaboration—note that activity-level elaboration (i.e., elaboration) is the one currently scored on the TTCT-figural.

Local independence: A concept which refers to the condition that the responses given to the items on a test be independent of one another (i.e., there is no correlation among the item scores after the latent trait measured by the items is controlled).

Unidimensionality: A concept which refers to the condition that a single variable be measured at a time (i.e., responses given to the items on a test are affected by one construct).

Differential item functioning: The loss of item fairness across groups. Differential item functioning addresses whether an item behaves differently for students in different groups with respect to gender, socioeconomic status, ethnicity, and etc.

Invariant measurement: Invariant measurement refers to the stability of item difficulty estimations across samples (i.e., sample-independent item calibration) or the stability of the trait level estimations of examinees across tests (i.e., item-independent measurements of the trait levels of examinees).

Variable map: A visual that shows the difficulty levels of the items on a test and the trait levels of examinees taking the test on the same continuum using the same metric.

The Torrance Tests of Creative Thinking, Figural Forms

The TTCT-figural are largely nonverbal tests that primarily require examinees to be involved in divergent thinking processes—divergent thinking is the ability to think in different directions and generate alternative ideas to a situation, question, or problem (Guilford, 1959; Taylor, 1988). There are two parallel forms of the TTCT-figural: Form A and Form B. Both forms are composed of three subtests (i.e., activities): Picture Construction, Picture Completion, and Repeated Figures (Torrance, 2006a, 2006b). The Picture Construction activity contains one item and the Picture Completion activity contains 10 items in both forms, while the Repeated Figures activity is comprised of 30 items in Form A and 36 items in Form B.

Each individual item is a visual stimulus (i.e., shape) and requires examinees to create a figural response using the item as the starting point (Ball & Torrance, 1984). Examinees are also asked to give a title to each figural response in all three activities (Torrance, 2006a, 2006b). Most often a figural response is the creation of a complete picture through drawing, but an intact item

can also be a figural response if the examinee gives the intact item meaning by putting a title (Ball & Torrance, 1984).

The items in the TTCT-figural are utilized to measure 18 creative thinking skills (i.e., variables). Five of the variables are norm-referenced variables. The norm-referenced variables are fluency, originality, elaboration, abstractness of titles, and resistance to premature closure (Torrance, 2008). The rest of the 13 variables are criterion-referenced variables and referred to as the creative strengths (Torrance, 2008). The creative strengths encompass the following creative thinking skills: emotional expressiveness, storytelling articulateness, movement or action, expressiveness of titles, synthesis of incomplete figures, synthesis of lines or circles, unusual visualization, internal visualization, extending or breaking boundaries, humor, richness of imagery, colorfulness of imagery, and fantasy (Ball & Torrance, 1984).

Responses (i.e., drawings and titles) are interpreted differently and evaluated separately for each of these 18 variables so that a response given to an item receives several scores on different variables measured on that item (Ball & Torrance, 1984). The item in the Picture Construction activity is utilized to measure all the variables with the exception of fluency, resistance to premature closure, synthesis of incomplete figures, synthesis of lines or circles, and extending or breaking boundaries. The items in the Picture Completion activity are used to measure all the variables except synthesis of lines or circles and extending or breaking boundaries. The items in the Repeated Figures activity permit the assessment of all the variables except abstractness of titles, resistance to premature closure, and synthesis of incomplete figures.

When a response given to an item is utilized to measure fluency or originality, the item is scored in a dichotomous manner (Ball & Torrance, 1984): Zero means that the variable is not manifested in the response, and 1 indicates that the variable is manifested in the response. In the

same way, a response given to an item can be scored in a dichotomous manner for each of the 13 creative strengths, although the creative strengths are criterion-referenced variables, and each test item does not currently receive a score for a particular creative strength as it does for fluency and originality. Because one of the purposes of the current study was to examine the properties of the test items for each of the 13 creative strengths, the test items were scored in a dichotomous manner for each creative strength for analyses purposes: Zero indicated that the creative strength was not expressed in the response, and 1 meant that the creative strength was expressed in the response.

When a response given to an item is utilized to measure abstractness of titles or resistance to premature closure, the item is scored in a polytomous manner (Ball & Torrance, 1984). This type of scoring is employed because both abstractness of titles and resistance to premature closure can be manifested to different degrees in a response. When each of these two variables is scored, a score is given based on the quality of the response (Torrance, Ball, & Safter, 1992). For instance, the title of a drawing can express four potential levels of abstractness; therefore, when scoring is done for abstractness of titles, responses are scored on a 4-point scale (Torrance et al., 1992).

Similarly, elaboration is also scored in a polytomous manner; however, when scoring is done for elaboration, each item does not receive an elaboration score. Instead, elaboration is scored in a polytomous manner at the activity level (i.e., activity-level elaboration). Accordingly, an activity (i.e., subtest) is considered as one item, and a score based on the total number of details in the activity is given to the activity as the elaboration score (Ball & Torrance, 1984; Torrance et al., 1992). For instance, in Activity 1 in Form A, 1 point is given if there are 0-5 details, 2 points are given if there are 6-12 details, 3 points are given if there are 13-19 details, 4

points are given if there are 20-26 details, 5 points are given if there are 27-33 details, and 6 points are given if there are more than 33 details (Torrance et al., 1992).

It is, however, possible to utilize the number of details added to a response in Activity 2 and Activity 3 to evaluate and score each individual item in a polytomous manner for ability to elaborate. Because one of the purposes of the current study was to examine the test items for ability to elaborate, the individual test items were scored in a polytomous manner for this ability. This ability was called item-level elaboration. From this point on, item-level elaboration is referred to as *elaboration-I*, and activity-level elaboration is referred to as *elaboration*. It should be noted here that elaboration-I was fabricated for the current study to investigate the properties of the individual items in Activity 2 and Activity 3 regarding item-level elaboration—how to score an item for elaboration-I is explained in the method section.

Because the test items were scored in a dichotomous manner for fluency, originality, and each of the 13 creative strengths in the study, these variables will be referred to as *the dichotomously scored variables* in the following sections. Elaboration, elaboration-I, abstractness of titles, and resistance to premature closure, on the other hand, will be referred to as *the polytomously scored variables* due to the scoring method. The content of each activity in the TTCT-figural, the variables measured in the tests, and the scoring method for each variable will be explained in more detail later in the paper.

Statement of the Problem

The TTCT-figural were developed following the principles of classical test theory (CTT), and the CTT framework has been used to standardize the tests since their inception (Torrance, 1966, 1974, 1984, 1990, 1998, 2008). As a measurement theory, CTT guides researchers in developing reliable tests. The CTT framework emphasizes total test scores (Crocker & Algina,

2008) and requires that the standard error of measurement involved in the total scores of examinees on a test be as low as possible (Engelhard, 2013). Hence, for decades, researchers have used total test scores, standard errors of measurement, and reliability coefficients when evaluating the quality of the TTCT-figural (e.g., Aslan & Puccio, 2006; Clapham, 2004; Torrance, 1981) and when standardizing the tests (Torrance, 1966, 1974, 1984, 1990, 1998, 2008).

Over the years, researchers have used one other measurement theory, factor analysis, to examine the TTCT-figural (e.g., Clapham, 1998, 2004; Kim, 2006; Prieto et al., 2006). Because factor analysis primarily focuses on the relationship between a set of items and the latent trait measured by these items (Engelhard, 2013; Sick, 2011), factor analysis studies on the TTCT-figural explored the factorial structure of the tests.

Unconventionally, these factor analysis studies did not examine the individual items in the TTCT-figural for each variable. Instead, each variable was considered as one item, the total score on each individual variable was used as an item score, and analyses were conducted at the variable level (e.g., Clapham, 1998; Kim, 2006). Inconsistent factorial structures were identified in these factor analysis studies: One-factor, two-factor, three-factor, and four-factor models were reported in different studies (see Aliotti & Blanton, 1969; Heausler & Thompson, 1988; Kim, 2006; Prieto et al., 2006).

Factor analysis or another measurement theory, other than CTT, has never been used to standardize the TTCT-figural. Therefore, CTT has remained the primary framework to evaluate the quality of the tests and to standardize them since their inception. The CTT framework has provided valuable information on the TTCT-figural and guided researchers in improving the quality of the tests to a certain extent. However, the common use of CTT has led to certain

assumptions about the TTCT-figural as well as the test items which have yet to be supported by sound statistical evidence.

Seven Assumptions About the Torrance Tests of Creative Thinking, Figural Forms

A first assumption is that all the items in the TTCT-figural measure the variables without any issues. This indicates that all the items have a good fit. The first assumption means that a particular variable is the only factor affecting the responses given to the items regarding that variable and that the responses given to the items are independent of one another—the former refers to unidimensionality, and the latter refers to local independence. Consequently, each of the 18 variables is evaluated separately, and each individual item contributes equally to estimating the total test scores (Torrance, 2008). It should be noted here that a few factor analysis studies (e.g., Auzmendi, Villa, & Abedi 1996; Kim, 2006) found two factors loaded on the same variable (e.g., fluency and resistance to premature closure), indicating that two different factors affect the responses given to the items regarding that variable. However, it is worth noting one more time that analyses were not conducted at the item level in these factor analysis studies.

A second assumption is that all the items in the TTCT-figural are at the same difficulty level regarding a dichotomously scored variable. This assumption applies to all the dichotomously scored variables. As a consequence, the manifestation of a particular dichotomously scored variable in a response is given 1 point across the items for that variable (Ball & Torrance, 1984).

A third assumption is related to the polytomously scored variables scored at the item level. It is assumed that all the items in the TTCT-figural are equally difficult regarding abstractness of titles and resistance to premature closure. For each of these two variables, the same type of response is accepted as corresponding to the same trait level across the items.

Therefore, the same score is assigned to the same type of response across the items regarding each of these two variables (Ball & Torrance, 1984). Similarly, it is assumed that all the items in Activity 2 and Activity 3 are equally difficult regarding elaboration-I. Consequently, every detail added to a response has the same weight regardless of the item to which the detail is added, and the same number of details is accepted as corresponding to the same trait level across the items in these two activities (Ball & Torrance, 1984).

A fourth assumption is about item scaling, and it concerns the polytomously scored variables scored at the item level. It is assumed that the response options of the items regarding abstractness of titles and resistance to premature closure are distributed evenly on the latent trait continuum. This assumption implies that the difficulty level between any two subsequent response options is the same across the response options. As a consequence, the response options of the items increase by 1 unit regarding each of these two variables (Ball & Torrance, 1984). In a similar fashion, regarding elaboration-I, every additional detail added to a response given to an item in Activity 2 or Activity 3 is assumed to require the same amount of increase in the trait level (i.e., the difficulty level between any two subsequent details is the same across the details).

A fifth assumption concerns elaboration. The response options of the activities also increase by 1 unit in all three activities regarding elaboration (Torrance et al., 1992). However, due to the number of items in each activity, the activities are not assumed to be equally difficult for elaboration. For instance, in Activity 1 in Form A, 1 point is given if there are 0-5 details; in Activity 2 in Form A, 1 point is given if there are 0-8 details; and in Activity 3 in Form A, 1 point is given if there are 0-7 details (Ball & Torrance, 1984). Additionally, although the response options of the activities increase by 1 unit, the number of details that corresponds to a response option is not the same for each response option. For instance, in Activity 3 in Form A, 1

point is given if there are 0-7 details, 2 points are given if there are 8-16 details, 3 points are given if there are 17-27 details, 4 points are given if there are 28-37 details, 5 points are given if there are 38-47 details, and 6 points are given if there are more than 47 details (Torrance et al., 1992).

The scaling of the activities aims to make the activities equally difficult with respect to average difficulty so that the same type of response (e.g., receiving 1 point) is accepted as corresponding to the same trait level in all three activities. The scaling of the activities also intends to make the response options be based on equal intervals. However, the scaling for elaboration is based on CTT analyses; therefore, the scaling method may not make the activities equally difficult and the response options be based on equal intervals.

A sixth assumption is about differential item functioning. The items in the TTCT-figural are assumed to behave the same for each gender and not to have significantly different difficulty levels for examinees of each gender. This assumption means that none of the test items favors examinees of one gender over examinees of the opposite gender. Therefore, all the test items are scored the same way for male and female examinees at all ages (Ball & Torrance, 1984).

A seventh assumption is about the standard error of measurement, and it concerns the norm-referenced variables (fluency, originality, elaboration, abstractness of titles, and resistance to premature closure). It is assumed that the standard error of measurement is the same for all scores on a particular norm-referenced variable and that all scores on that variable are equally reliable (e.g., Aslan & Puccio, 2006; Clapham, 2004; Prieto et al., 2006; Torrance, 2008). As a consequence, one standard error of measurement and one reliability coefficient are estimated for all scores on a norm-referenced variable.

The seventh assumption indicates that each individual item in the TTCT-figural provides the same amount of information on a particular norm-referenced variable and that the amount of information provided by a particular item on this variable is at the same level along the latent trait continuum. The seventh assumption also indicates that each of the three activities provides the same amount of information on elaboration and that the amount of information provided by a particular activity on elaboration is at the same level along the latent trait continuum.

These seven assumptions have not yet been examined. Due to the absence of supporting evidence, these assumptions raise methodological concerns. Additionally, they may affect the interpretations of the test scores because the measurement scale of a test impacts linear estimations, such as standard deviations and correlations (Embretson & Reise, 2000), and these linear estimations affect the computations of the standard error of measurement and reliability as well as validity coefficients. Thus, the items in the TTCT-figural must be analyzed using an appropriate method to examine whether these assumptions are supported by statistical evidence.

Limitations of Classical Test Theory

Although CTT focuses on the total test score, it is possible to examine certain properties of the items on a test using the CTT framework. However, the CTT framework has certain limitations. In CTT, the difficulty level of a dichotomously scored item is the proportion of examinees answering the item correctly (Bond & Fox, 2001; Crocker & Algina, 2008; De Champlain, 2010), but because item difficulty is estimated in a linear fashion and is not based on a true interval scale, item difficulties cannot be calibrated accurately and compared appropriately (DeMars, 2010).

Similarly, an average item score is used to estimate the average difficulty level of a polytomously scored item (DeMars, 2010), but an average item score does not offer any

information about the distributions and difficulty levels of the response options (i.e., item scaling). Therefore, under the CTT framework, researchers assume that the response options of the polytomously scored items on a test are based on equal intervals and that these intervals are stable across the items (Bond & Fox, 2001).

Another limitation concerns the sample-dependent nature of the item parameters estimated under CTT. Both the difficulty levels of dichotomously scored items and the average item scores of polytomously scored items are sample-dependent: They change as the characteristics of a sample used for item calibration change (Embretson & Reise, 2000).

One other limitation is related to item fit. Because CTT focuses on total scores on a test, the CTT framework offers no method to assess item fit. In other words, unidimensionality and local independence cannot be addressed under the CTT framework.

Finally, the CTT framework yields only one standard error of measurement, which is the same for all scores on a variable for a given age or grade on the same test (Crocker & Algina, 2008). In other words, all scores on a particular variable are accepted to be equally reliable under CTT. Due to these limitations, the seven assumptions about the TTCT-figural cannot be thoroughly examined through the CTT framework.

Limitations of Factor Analysis

As an alternative measurement theory, factor analysis can address the assumption about item fit (i.e., unidimensionality), but all seven assumptions about the TTCT-figural cannot be studied thoroughly through factor analysis either. Different from CTT, factor analysis focuses on the individual items on a test and offers information on each item; however, factor analysis has certain limitations as well. As in CTT, item parameters are estimated in a linear fashion in factor

analysis (Reise, Keith, & Robin, 1993), and analyses deliver sample-dependent item parameters (Ewing, Salzberger, & Rudolf, 2005).

Another limitation concerns item difficulty. The concept of item difficulty has no clear equivalent in factor analysis (Meade & Lautenschlager, 2004). Therefore, factor analysis cannot appropriately address the difficulty levels of dichotomously scored items and the difficulty levels of the response options of polytomously scored items (i.e., item scaling).

Finally, factor analysis delivers one standard error of measurement for all scores on a test (Ewing et al., 2005). Therefore, all scores on the same test are accepted to involve the same amount of standard error of measurement and to be equally reliable under factor analysis.

It should be noted here that factor analysis can address whether the items on a test measure only one latent variable. However, if the items on the test vary in difficulty and if the researcher neglects this variation, the researcher likely interprets the results falsely and thus identifies misleading factorial structures (Sick, 2011).

Rasch Measurement Theory

Due to the limitations of CTT and factor analysis, another measurement theory that can thoroughly examine the seven assumptions about the TTCT-figural must be utilized. Rasch measurement theory (Rasch, 1960) appears to be a perfect candidate for this purpose. Rasch measurement theory can not only examine all seven assumptions about the TTCT-figural but also offer further information on the tests.

It should be noted that dichotomous item response theory models (e.g., the two-parameter logistic model) and polytomous item response theory models (e.g., the graded response model) can also be used to address the seven assumptions about the TTCT-figural. However, Rasch measurement theory analyses provide stronger justification for the properties of the items on a

test (Embretson & Reise, 2000). Additionally, only Rasch measurement theory delivers the desired level of invariant calibration of the items and invariant measurement of the trait levels of examinees (Bond & Fox, 2001; Embretson & Reise, 2000; Engelhard, 2013).

In Rasch measurement theory, a person's latent trait level is related to his or her responses given to the items on a test (Engelhard, 2013). Rasch measurement theory focuses on the responses given to each individual item on a test separately and takes the difficulty levels of the items into consideration. Under the Rasch measurement theory framework, there is a nonlinear relation between the trait level of an examinee and item as well as test performance (Bond & Fox, 2001). Using this nonlinear relation, Rasch measurement theory answers the following question: "[W]hat trait level is most likely to explain the person's responses?" (Embretson & Reise, 2000, p. 54).

Under Rasch measurement theory, the main objective is to develop *ideal-type scales*. Ideal-type scales are instruments (e.g., tests or questionnaires) that deliver not only reliable test scores but also invariant calibration of items and invariant measurement of the trait levels of examinees (Engelhard, 2013). To develop an ideal-type scale, Rasch measurement theory posits four requirements.

First, it is required that an ideal-type scale deliver invariant (i.e., item-independent) measurement of the latent trait levels of examinees (Bond & Fox, 2001). This implies that no matter to which items on a test an examinee has responded, the examinee's latent trait level should be found to be the same upon measurement.

The second requirement states that an ideal-type scale yields invariant (i.e., sample-independent) item calibration (Bond & Fox, 2001). This means that no matter whose responses are used to estimate item parameters (e.g., item difficulty), item parameters should be found to

be at the same level (Engelhard, 2013). The second requirement also means that examinees with the same trait level cannot have significantly different performances on the items in an ideal-type scale; otherwise, the items are considered to be biased against at least one group of examinees (Engelhard, 2013).

According to the third requirement, the items in an ideal-type scale have the same level of item discrimination. The third requirement indicates that in an ideal-type scale, not only does an examinee with a higher trait level always have a better chance of answering an item correctly than does an examinee with a lower trait level, but also all examinees have a better chance of correctly answering an easy item than a difficult item (Engelhard, 2013). These indications are for items scored in a dichotomous manner. In the same fashion, in an ideal-type scale containing polytomously scored items, not only does an examinee with a higher trait level always respond to a higher response option than does an examinee with a lower trait level, but also all examinees have a better chance of responding to a lower response option than to a higher response option.

The fourth requirement concerns item fit. The final requirement states that all the items in an ideal-type scale have a good item fit. Good item fit implies that scores given to examinees based on their responses are indications of one latent trait and that the responses given to the items are independent of one another (Bond & Fox, 2001).

Item parameters. Rasch measurement theory offers several item parameters. These parameters are item fit, item difficulty, category thresholds, item information, and item reliability index—item parameters will be explained in more detail in the method section.

Item fit. In Rasch measurement theory, item fit is an indication of the quality of an item (Bond & Fox, 2001; Engelhard, 2013). Researchers utilize item fit statistics to evaluate whether an item has any issues that may cause inadequate measurements of the trait levels of examinees.

Fit statistics show whether the items on a test have a good fit or misfit. Good item fit implies that the responses given to an item are affected by one factor and independent of each other (Bond & Fox, 2001). Misfit, on the other hand, denotes either a second factor impacting the student responses or the predictability of the responses given to an item.

Item difficulty. Item difficulty shows how difficult it is to correctly answer a dichotomously scored item (Bond & Fox, 2001). For a polytomously scored item, average item difficulty is estimated (under the rating scale model). In simple terms, average item difficulty shows how difficult a polytomously scored item is (Bond & Fox, 2001). Item difficulties are expressed in logits which are units of measurement based on an interval scale. Item difficulties are constrained to a particular range such as between -3 and +3, with a mean value of 0. Difficulty levels increase as the value of item difficulty goes from negative to positive.

Category thresholds. Category thresholds are estimated for polytomously scored items. Category thresholds are points on the latent trait continuum where two subsequent response options have the same probability of being responded to by examinees (Bond & Fox, 2001; Engelhard, 2013). For instance, for a rating scale item (i.e., Likert-type item) with five response options, there are four thresholds—each threshold is between any two subsequent response options. Category thresholds indicate how difficult each response option of a polytomously scored item is. Therefore, category thresholds show whether the response options of a polytomously scored item are equally distributed (Engelhard, 2013).

Item information. In Rasch measurement theory, item information is tied to reliability (Baker, 2001; Cappelleri, Lundy, & Hays, 2014). The amount of information an item provides on the measured variable is estimated through the item information function. The item information function allows researchers to estimate how much standard error of measurement is involved in

an item at a particular trait level. The amount of standard error of measurement involved in an item increases as the amount of information provided by the item decreases (Baker, 2001).

The amount of information provided by the individual items on a test at a particular trait level can be totaled to estimate the amount of information provided by the test at that trait level (Baker, 2001; Embretson & Reise, 2000). Therefore, the information function of a test indicates how much standard error of measurement is involved in the test scores at each point on the latent trait continuum.

***Item reliability index*.** Item reliability index is used to assess whether the items on a test can be calibrated in a sample-independent (i.e., invariant) manner (Bond & Fox, 2001). Item reliability index indicates whether the item parameters (e.g., item difficulty) will remain at the same level if the instrument is administered to another group of examinees (Wright & Masters, 1982).

Advantages of Rasch measurement theory. The features of Rasch measurement theory offer major advantages over CTT and factor analysis. In Rasch measurement theory, the difficulty level of a dichotomously or polytomously scored item and the category thresholds of a polytomously scored item are estimated through a natural logarithmic function (Engelhard, 2013). This logarithmic function transfers raw item scores, which are based on an ordinal scale, into logits, which are based on an interval scale (Bond & Fox, 2001; Engelhard, 2013). Because item difficulty is expressed in logits and based on an interval scale, items can be compared accurately in terms of difficulty (Bond & Fox, 2001).

Another advantage is that Rasch measurement theory allows researchers to estimate item fit indices for each individual item on a test. Under Rasch measurement theory, item fit is used to evaluate whether the variable measured by a set of items is the only factor affecting the

responses given to the items and whether the responses given to the items on a test are independent of each other (Bond & Fox, 2001).

One additional advantage is that Rasch measurement theory analyses yield more than one standard error of measurement. In other words, under Rasch measurement theory, not all scores on a test are accepted to be equally reliable. Researchers use item information and test information functions to estimate the standard error of measurement involved in an item and a test score at a particular trait level (Baker, 2001; Cappelleri et al., 2014).

Finally, Rasch measurement theory analyses show whether a test delivers invariant calibration of the items and invariant measurement of the trait levels of examinees (Engelhard, 2013). In other words, Rasch measurement theory addresses whether the item difficulties are estimated in a sample-independent manner and whether the trait levels of examinees are determined in an item-independent fashion.

Limitations of Rasch measurement theory. Rasch measurement theory has a few limitations. Rasch measurement theory requires researchers to possess some degree of knowledge of statistics due to the complex computations (Hambleton & Jones, 1993). The complexity of the computations of item and person parameters may be confusing when these parameters are interpreted.

Another limitation is related to the sample size. Relatively large samples are needed to obtain accurate item parameters when Rasch measurement theory is used for item analyses (Hambleton & Jones, 1993).

One other limitation is about the estimations of item difficulty. The Rasch model does not yield as precise estimations of item difficulty as the two-parameter logistic model does for the same data (Pelton, 2002).

Finally, the two-parameter logistic model and the three-parameter logistic model always fit the same data better than the Rasch model (Ghaemi, 2012). This is because the two-parameter logistic model and the three-parameter logistic model take additional parameters—item discrimination and guessing parameter respectively—into consideration. In Rasch measurement theory, on the other hand, all the items are projected to be equally discriminating, and there is no guessing parameter.

Purpose of the Present Study

The purpose of the present study was to conduct Rasch measurement theory analyses for the items regarding each of the 18 variables measured in the TTCT-figural Form A as well as elaboration-I to examine the assumptions about item fit, item difficulty, item scaling, differential item functioning, and item information. Additionally, the study examined whether the TTCT-figural Form A delivers invariant calibration of the items as well as invariant measurement of the creative thinking skill levels of second grade students in Turkey.

Analyses were conducted separately for each of the 18 variables measured in the test as well as for elaboration-I. The test items were analyzed for fluency, originality, abstractness of titles, resistance to premature closure, each of the 13 creative strengths, and elaboration-I. Because elaboration is scored at the activity level, analyses were conducted at the activity level for elaboration. The research questions of the current study were as follows:

Question 1: How good is item fit for each individual variable measured in the TTCT-figural Form A as well as for elaboration-I?

1. Is a particular variable the only factor affecting the responses given to the items regarding that variable?

2. Are the responses given to the items independent of each other regarding a particular variable?

Question 2: What are the levels of item difficulty for each individual variable, including elaboration-I, measured in the TTCT-figural Form A?

1. What are the levels of item difficulty for the dichotomously scored variables and average item (or activity for elaboration) difficulty for the polytomously scored variables? For each dichotomously scored variable, is it equally difficult to manifest a particular variable in a response across the items? Are all the items (or activities) at the same average difficulty level regarding a polytomously scored variable?
2. Are the response options of the items (or activities) distributed evenly on the latent trait continuum regarding polytomously scored variables? Regarding elaboration-I, does every additional detail added to response given to an item in Activity 2 and Activity 3 require the same amount of increase in the trait level?
3. What are the difficulty levels of the items (or activities) for each gender regarding a particular variable? Does each test item (or activity) behave the same for each gender regarding a particular variable?

Question 3: How much standard error of measurement is involved in scores on each norm-referenced variable on the latent trait continuum?

1. Is the standard error of measurement involved in each individual item (or activity for elaboration) the same across the items for each of the norm-referenced variables?

2. Is the standard error of measurement involved in each individual item (or each of the three activities) at the same level along the latent trait continuum?

Question 4: Does the TTCT-figural Form A deliver invariant calibration of the items and invariant measurement of the creative thinking skill levels of second grade students?

1. Are the difficulty levels of the items (or activities for elaboration) determined in a sample-independent manner for a particular variable, so the order of the items (or activities) with regard to their difficulty levels will remain the same if another group of second graders are administered the test?
2. Are the creative thinking skill levels of examinees measured in an item-independent manner by the TTCT-figural Form A so that the trait levels of second grade students would remain at the same level if they were given another set of items measuring the same construct (e.g., TTCT-figural From B)?

Significance of the Present Study

Because the Rasch measurement theory framework has not yet been employed to examine the seven assumptions about the TTCT-figural, currently there is no sound statistical evidence supporting these assumptions. Additionally, there is no statistical evidence justifying that the TTCT-figural deliver invariant calibration of the items as well as invariant measurement of the creative thinking skill levels of examinees. Evidently, there is a need for research that addresses these issues. Regarding its focus, the current study is the first of its kind.

With its focus on the analyses of item fit, item difficulty, item scaling, differential item functioning, item information, and invariant calibration of the items as well as invariant measurement of the creative thinking skill levels of examinees, the current study provided certain implications for the TTCT-figural Form A. Analyses of item fit showed whether an item or

activity had any issues that could cause inadequate measurements of the creative thinking skill levels of second grade students. Specifically, analyses of item fit examined construct validity of the TTCT-figural Form A (i.e., unidimensionality) and item independence (i.e., local independence) for each individual variable measured in the test as well as for elaboration-I.

Analyses of item difficulty and item scaling offered suggestions for the scoring method of the items regarding the variables measured in the test. Additionally, item difficulty analyses detected the item(s) that behaved differently for each gender regarding a particular variable.

Analyses of item information revealed whether the items measured the creative thinking skills with equal precision at all trait levels regarding the norm-referenced variables. In other words, analyses of item information examined whether the standard error of measurement was the same for all scores on a norm-referenced variable.

Finally, item analyses addressed invariant measurement. Rasch measurement theory analyses showed whether the test items could be calibrated in a sample-independent manner and whether the creative thinking skill levels of second grade students could be measured in an item-independent manner by the TTCT-figural Form A.

CHAPTER 2

LITERATURE REVIEW

The literature review is composed of three major parts in pursuit of familiarizing readers with the TTCT-figural, highlighting gaps in the literature, and showing applications of Rasch measurement theory in creativity testing. First, the content of the TTCT-figural, the variables measured in the tests, and the scoring method for each variable will be described. Second, studies that examined the reliability, validity, and fairness of scores on the TTCT-figural will be summarized. Finally, studies using Rasch measurement theory to examine instruments that measure the potential for creativity will be discussed.

The Torrance Tests of Creative Thinking, Figural Forms

Torrance started his academic research on creativity at the University of Minnesota in 1958 (Torrance, 1988). Within few years, Torrance and his colleagues developed several creativity tasks, including verbal and nonverbal tasks, and published one verbal test and one nonverbal test (Goldman, 1965; Torrance, 1968). The nonverbal test was called the Minnesota Tests of Creative Thinking, Nonverbal Form. Like the TTCT-figural, the nonverbal form of the Minnesota Tests of Creative Thinking was composed of three subtests, but the majority of the test items on the test were different from the items in the TTCT-figural.

In 1966, Torrance left the University of Minnesota and began his academic career at the University of Georgia. Soon after he moved to the University of Georgia, Torrance started to make revisions in the Minnesota Tests of Creative Thinking. These revisions primarily involved

developing new test items. After reviewing a number of studies on the Minnesota Tests of Creative Thinking, he completed the revisions and published the TTCT-figural (Torrance, 1966).

The TTCT-figural (Torrance, 1966) come in two parallel forms: Form A and Form B. Each form is comprised of three subtests (i.e., activities). These activities are Picture Construction, Picture Completion, and Repeated Figures (Torrance, 2006a, 2006b). The Picture Construction activity contains one item and the Picture Completion activity contains 10 items in both forms, while the Repeated Figures activity is comprised of 30 items in Form A and 36 items in Form B. Each individual item is a visual stimulus in a particular shape.

In each activity, examinees are given 10 minutes to respond and asked to create a figural response for each item in the activity. When a figural response is created, the stimulus must be used as an integral part of the response. Examinees are also asked to give a title to each figural response in all three activities (Torrance, 2006a, 2006b). Examinees are encouraged to work on as many items as they can, but they are not required to respond to all the items. The TTCT-figural can be administered individually or in groups (Torrance, 1966). The content of the TTCT-figural, the variables measured in the tests, and the scoring method for each variable will be explained below.

Activity 1: Picture Construction

The Picture Construction activity was developed by Torrance (1962). In this activity, there is only one item which is at the center of the page. In Form A, the item looks like an egg; in Form B, the item looks like a pickle. The Picture Construction activity was developed to stimulate "the tendency toward finding a purpose for something that has no definite purpose" (Torrance, 2000, p. 4). With the exception of fluency, resistance to premature closure, synthesis of incomplete figures, synthesis of lines or circles, and extending or breaking boundaries, all the

variables are measured in Activity 1. Examinees receive the following instruction before they start working on the activity in both forms:

> On the opposite page is a curved shape. Think of a picture or an object which you can draw with this shape as a part. Try to think of a picture that no one else will think of. Keep adding new ideas to your first idea to make it tell as interesting and as exciting a story as you can. When you have completed your picture, think up a name or title for it and write it at the bottom of the page in the space provided. Make your title as clever and unusual as possible. Use it to help tell your story. (Torrance, 1966, 1984, p. 2)

Activity 2: Picture Completion

The second activity, Picture Completion, was adapted from the Franck Drawing Completion Test (Franck & Rosen, 1949). The Franck Drawing Completion Test consisted of 36 incomplete figures which were used to assess whether the responses of an examinee reflected a masculine or feminine character (Harkey, 1982). Torrance decided to utilize the responses given to incomplete figures to measure creative thinking skills. For this purpose, Torrance developed 20 incomplete figures—10 figures for each form—with the assistance of Elizabeth Kennedy (Torrance, 1974). These incomplete figures were inspired from complete pictures drawn by examinees in a previously conducted study (Torrance, 1974).

In Activity 2, there are 10 items (i.e., incomplete figures) which are displayed on two pages. The page on the left has four and the page on the right has six incomplete figures. There are two figures in each row, and each figure is in a square. The Picture Completion activity "creates tension in the beholder, who must control this tension long enough to make the mental leap necessary to get away from the obvious and commonplace" (Torrance, 2000, p. 4). With the exception of synthesis of lines or circles and extending or breaking boundaries, all the variables

are measured in Activity 2. Examinees receive the following instruction before they start working on the activity in both forms:

> By adding lines to the incomplete figures on this and the next page, you can sketch some interesting objects or pictures. Again, try to think of some picture or object that no one else will think of. Try to make it tell as complete and as interesting a story as you can by adding to and building up your first idea. Make up an interesting title for each of your drawings and write it at the bottom of each block next to the number of the figure. (Torrance, 1966, 1984, p. 4)

Activity 3: Repeated Figures

The Repeated Figures activity was one of the first activities developed by Torrance and his colleagues when Torrance was at the University of Minnesota (Torrance, 1979). Activity 3 is composed of 30 pairs of parallel lines in Form A and 36 circles in Form B. The items in the Repeated Figures activity are not separated from each other by lines.

The Parallel Lines task is displayed on three pages. On the first page, there are six pairs of parallel lines, with three pairs in each row; on the second and third page, there are 12 pairs of parallel lines, with three pairs in each row. The pairs of parallel lines differ from each other in terms of the distance between the two parallel lines in each pair. There are three types of parallel line pairs. The pairs of parallel lines in the first row on the first page are approximately 0.25 inches apart. The pairs of parallel lines on the second page and in the second row on the first page are approximately 0.75 inches apart. The pairs of parallel lines on the third page are approximately 0.50 inches apart.

The Circles task is displayed on two pages. On the first page, there are six circles, with three circles in each row. On the second page, there are 30 circles, with five circles in each row. All circles in the Circles task are at the same size.

The Repeated Figures activity requires "an ability to return to the same stimulus again and again, perceiving it differently each time, [and] disrupting structure in order to create something new" (Torrance, 2000, p. 4). With the exception of abstractness of titles, resistance to premature closure, and synthesis of incomplete figures, all the variables are measured in Activity 3. Examinees receive the following instruction before they start working on the activity in Form A:

> In ten minutes see how many objects or pictures you can make from the pairs of straight lines below and on the next two pages. The pairs of straight lines should be the main part of whatever you make. With pencil or crayon add lines to the pairs of straight lines to complete your picture. You can place marks between the lines, on the lines, and outside the lines—wherever you want to in order to make your picture. Try to think of things that no one will think of. Make as many different pictures or objects as you can and put as many ideas as you can in each one. Make them tell as complete and as interesting a story as you can. Add names or titles in the spaces provided. (Torrance, 1966, 1984, p. 6)

In Form B, the phrase *pairs of straight lines* is replaced with the word *circles*, and examinees are told to work on two pages (Torrance, 2006b).

Variables Measured in the Torrance Tests of Creative Thinking, Figural Forms

Currently, 18 different variables are measured in the TTCT-figural: fluency, originality, elaboration, abstractness of titles, resistance to premature closure, emotional expressiveness, storytelling articulateness, movement or action, expressiveness of titles, synthesis of incomplete

figures, synthesis of lines or circles, unusual visualization, internal visualization, extending or breaking boundaries, humor, richness of imagery, colorfulness of imagery, and fantasy (Torrance, 2008). These variables will be explained below.

Fluency. Research on fluency goes back to the late 1930s and early 1940s (see Carroll, 1941; Johnson & Reynolds, 1941; Thurstone, 1938). The objective of the initial research was verbal fluency, and fluency was defined "as the ability to produce words in accordance with some restriction" (Taylor, 1947, p. 239). Guilford (1962) recognized the initial research on fluency, and for creativity research purposes, he described fluency as the quantity of responses and flow of ideas. Torrance (1966, 1979) embraced this description when developing his tests and defined fluency as the ability to produce a large number of ideas.

In the TTCT-figural, fluency is the ability to create a meaningful figural response using a particular item as the starting point. When responses are utilized to measure fluency, the items are scored in a dichotomous manner: Zero means that the item is not used to create a meaningful response, and 1 indicates that the item is used.

Fluency is the first variable to be scored in the TTCT-figural, and it is measured in Activity 2 and Activity 3. The item in Activity 1 is not scored on fluency, but the rater must decide whether the item is used (Torrance et al., 1992). If an item is not used to create a meaningful figural response, the item is not scored on any other variables, so the item receives 0 points in total (Torrance et al., 1992).

It is sometimes observed that an examinee combines two or more items into one drawing. A combination of items can be observed in both Activity 2 and Activity 3. The use of combinations affects the fluency scoring. When two or more items are combined in Activity 2, 1 fluency point is given to each item used in the combination (Torrance et al., 1992). Because it is

relatively common to combine items in Activity 3, scoring for fluency is different. When two or more items are combined in Activity 3, each combined response is counted as one response for fluency and receives 1 point (Ball & Torrance, 1984).

Originality. Originality has long been of interest in creativity research (see Wilson et al., 1953; Wilson, Guilford, Christensen, & Lewis, 1954). Originality is the production of unusual ideas (Guilford, 1962), and it refers to the rarity of an idea (Torrance, 1962, 1979). Originality is usually highly correlated with fluency (Torrance, 2008). Therefore, the possibility of generating original ideas increases with high fluency. High fluency, however, does not necessarily guarantee high originality (Torrance, 1979). In addition, originality predicts creative behavior better than fluency (Torrance, 1972).

In the TTCT-figural, originality is the ability to create an original figural response using a particular item (Torrance, 1962, 1979). To help raters score originality easily and objectively, Torrance (1966, 1974, 1984) identified the common (i.e., unoriginal) responses given to each item and put them in the scoring manual. These common responses were determined using the responses given by 500 randomly selected examinees (Chase, 1985). For a response to be considered original, the response should not be among the common responses (Ball & Torrance, 1984). When responses are utilized to measure originality, the items are scored in a dichotomous manner: Zero means that the item is not used to create an original response, and 1 indicates that the response given to the item is original. Originality is measured in all three activities.

When originality is scored, a combination of items is given bonus points because combining items into one drawing is an indication of an original way of thinking (Torrance, 1979). These bonus points are given to an examinee in addition to his or her originality score. In Activity 2, a combination of two items receives 3 bonus points, a combination of three items

receives 4 bonus points, a combination of four items receives 5 bonus points, and so on (Torrance et al., 1992). In Activity 3, scoring for bonus points is different because a combination of items is more common to observe (Ball & Torrance, 1984). If two items are combined, 1 bonus point is given; if 3-5 items are combined, 2 bonus points are given; if 6-10 items are combined, 3 bonus points are given; if 11-15 items are combined, 4 bonus points are given; and if more than 15 items are combined, 5 bonus points are given (Ball & Torrance, 1984).

Elaboration. While examining planning skills, Berger, Guilford, and Christensen (1957) identified an ability called elaboration. Since then, elaboration has been recognized as a creative thinking skill. In creativity, elaboration is defined as the number of supporting ideas added to the initial idea (Guilford, 1962; Torrance, 1962, 1979). When developing his tests, Torrance (1966) recognized this ability and included in his tests.

In the TTCT-figural, elaboration is the ability to add details to a response. To be counted for elaboration, those details should be beyond the minimum necessary form of details to draw the objects or the characters in the response (Torrance et al., 1992). For instance, when drawing a face, minimum necessary details would be the eyes, nose, and mouth; dimples, wrinkles, or earrings, would be extra details. A descriptive title can also be an indication of elaboration. Elaboration is counted at the item level, which is item-level elaboration (i.e., elaboration-I), but scored at the activity level, which is activity-level elaboration (i.e., elaboration).

Elaboration is scored on a 6-point scale. When scoring is done for elaboration, the rater first counts the number of details added to each response (i.e., drawing and title) in the activity and totals the number of details in the response or across the responses in the activity. Afterwards, the rater gives a score to the activity based on the total number of details in the activity. For instance, in Activity 1 on Form A, 1 point is given if there are 0-5 details, 2 points

are given if there are 6-12 details, 3 points are given if there are 13-19 details, 4 points are given if there are 20-26 details, 5 points are given if there are 27-33 details, and 6 points are given if there are more than 33 details (Torrance et al., 1992). Activity 2 and Activity 3 also are scored on a 6-point scale, but the number of details that corresponds to each point is slightly different in these two activities. For instance, in Activity 2 in Form A, 1 point is given if there are 0-8 details, while in Activity 3 in Form A, 1 point is given if there are 0-7 details (Ball & Torrance, 1984). This type of scoring is employed because there are different numbers of items in each activity, and therefore, a different range of possible responses.

Abstractness of titles. Abstractness of titles is an indicator of an ability which Torrance (1979, 2000) called *highlighting the essence*. Highlighting the essence involves abstract thinking, synthesis, and filtering ideas (Torrance, 1979). In the TTCT-figural, abstractness of titles is the ability to give an abstract title to a drawing (Ball & Torrance, 1984). Such a title captures the essence of the idea in the response.

When titles are utilized to measure abstractness of thinking, the items are scored on a 4-point scale. If the title of a drawing is not an abstract title and just tells what is in the drawing (e.g., ball), the item receives 0 points (Torrance, 1979). If the title contains a descriptor at a concrete level (e.g., colorful ball), the item receives 1 point (Torrance, 1962). If the title moves away from the concrete description of the drawing (e.g., I like to play soccer), the item receives 2 points (Torrance et al., 1992). Finally, if the title describes the drawing in an abstract form (e.g., fun times), the item receives 3 points (Torrance, 1979). Abstractness of titles is measured in Activity 1 and Activity 2.

Resistance to premature closure. Before the TTCT-figural were developed, closure was already an interest of Gestalt psychology (see Schoenfeld, 1941). According to Schoenfeld

(1941), closure occurs in perception. Accordingly, when the human mind is stimulated through an incomplete object or image, it tends to complete the missing parts of this object or image. Based on this view, Torrance (1979) described closure as tension an examinee feels when facing with an incomplete figure.

When working on the incomplete figures in Activity 2, examinees have the urge to close the open part of each incomplete figure as quickly as possible (Torrance, 1979). However, creativity requires examinees to delay this closure and to consider alternative perspectives before making a decision (Torrance, 2000). According to Ball and Torrance (1984), resistance to premature closure is the ability to resist or delay closing the open part of an incomplete figure to create an unconventional response.

Because Activity 2 is the only one with incomplete figures, resistance to premature closure is measured only in Activity 2. When responses are utilized to measure resistance to premature closure, the items are scored on a 3-point scale. Zero indicates the lowest resistance to premature closure, as when the examinee closes the figure with 1-3 straight or simple curved lines as quickly as possible; 2 indicates the highest resistance to premature closure, as when the examinee leaves the figure open or closes the figure incidentally; and 1 means that the response is somewhere in between the lowest and the highest resistance to premature closure (Torrance et al., 1992).

Creative strengths. In 1982, Torrance decided to include 13 creative strengths in the TTCT-figural due to their importance in creative thinking (Torrance, 1984). Torrance added these creative strengths as criterion-referenced variables so that if a creative strength is manifested once or twice in the entire test, the creative strength is given 1 point; and if the

creative strength is manifested more than two times, the creative strength is given 2 points (Torrance et al., 1992).

Richness of imagery, however, has different requirements because the TTCT-figural are figural tests. If richness of imagery is manifested four or five times in the entire test, it receives 1 point; and if richness of imagery is manifested more than five times, it receives 2 points (Ball & Torrance, 1984).

No matter how many expressions of the same creative strength are present in a response, the creative strength is scored just once in each item. For instance, if humor is expressed in four different ways in one item but not expressed in the other items, humor receives only 1 point.

The manifestation of a particular creative strength is not scored in each item as are fluency and originality; however, a response given to an item can be evaluated and scored in a dichotomous manner for a particular creative strength. For analyses purposes, the test items were scored in a dichotomous manner for each of the 13 creative strengths in the current study: Zero meant that the creative strength was not manifested in the response, and 1 indicated that the creative strength was manifested in the response. Each creative strength will be described below.

Emotional expressiveness. Emotional factors are important for creative performance (Torrance, 1979). Therefore, creative individuals are expected to reflect emotions when being involved in a creativity process. Emotional expressiveness is the ability to show some kind of emotion, such as fear and happiness, in a response (Ball & Torrance, 1984). Emotional expressiveness can be observed in a drawing or the title of a drawing (Torrance, 2000). Emotional expressiveness is measured in all three activities.

Storytelling articulateness. Storytelling articulateness has connections with conveying ideas effectively (Torrance, 1979). This creative strength is the ability to tell a story through an

interaction among the characters in a response or the creation of an environment (Torrance et al., 1992). The title of a drawing can also indicate a story (Ball & Torrance, 1984). Storytelling articulateness is measured in all three activities.

Movement or action. According to Torrance (1979), creative individuals, especially children, tend to use kinesthetic and auditory senses for creativity. Therefore, creative individuals are expected to reflect some sort of action when being involved in a creativity process. Movement or action is the ability to create a drawing that involves people, animals, vehicles, or other objects that move (Torrance, 2000). The title of a drawing can also indicate movement or action (Torrance et al., 1992). Movement or action is measured in all three activities.

Expressiveness of titles. Expressiveness of titles is an indication of transforming ideas into emotions and then expressing them in words (Torrance et al., 1992). This creative strength is the ability to give a drawing a title that induces feelings or that has rich descriptions (Torrance et al., 1992). If a drawing is difficult to comprehend without its title, the title receives a point for this creative strength (Ball & Torrance, 1984). Expressiveness of titles is measured in all three activities.

Synthesis of incomplete figures. Combination of items into one drawing has connections with seeing relationships among unrelated objects or ideas (Torrance, 1979; Torrance et al., 1992). Synthesis of incomplete figures is the ability to break boundaries and combine two or more incomplete figures into one drawing (Torrance et al., 1992). A title can also indicate a synthesis of two or more figures. Synthesis of incomplete figures is measured in only Activity 2.

Synthesis of lines or circles. As similar with synthesis of incomplete figures, synthesis of lines or circles is related to seeing connections among unrelated objects or ideas (Torrance et al.,

1992). This creative strengths is the ability to combine two or more items (i.e., parallel lines or circles) in Activity 3 into one drawing (Torrance et al., 1992). A title can also imply that several lines or circles are combined. Synthesis of lines or circles is measured only in Activity 3.

Unusual visualization. Unusual visualization is an indication of having an unusual perspective towards the world (Torrance & Safter, 1999). This creative strength is the ability to create a drawing that includes an object or a character which is portrayed in an unusual way such as from above, underneath, at an unusual angle, or an unusual distance (Torrance et al., 1992). Unusual visualization is measured in all three activities.

Internal visualization. Creative individuals' points of view towards the world go beyond the exterior features of objects, and these individuals can focus on hidden problems (Torrance, 1979). Internal visualization is the ability to create a drawing that shows the interior of something, such as the interior of a car (Torrance, 1979), or that shows things that are inside other things such as fish in an aquarium (Torrance, 2000), or a cross-section. Internal visualization is measured in all three activities.

Extending or breaking boundaries. In order to synthesize unrelated objects or ideas and generate creative ideas, one has to extend or break the boundaries and move away from ordinary ideas. Extending or breaking boundaries is the ability to extend the boundaries of parallel lines vertically or horizontally, to give parallel lines or circles depth, or to extend circles in any direction (Torrance et al., 1992). Extending or breaking boundaries is measured in only Activity 3.

Humor. Humor is usually considered as a type of creativity, and it stimulates creative thinking (Torrance, 1979). By nature, humor involves some sort of originality and is an indicator of creative personality (Torrance, 1979). Humor is the ability to create a drawing or to give a

drawing a title that shows comedy, irony, silliness, sarcasm, or a similar concept (Torrance et al., 1992). Humor is measured in all three activities.

Richness of imagery. Ability to generate visually appealing ideas is crucial for creative performance in many fields (Torrance, 1979). This creative strength is the ability to create a drawing that displays vividness, liveliness, and intensity, such as creating a twist on a common object or an unusual depiction of an ordinary thing (Ball & Torrance, 1984). Some examples include a deflated football rather than a regular football, or the use of parallel lines to draw a detailed dead tree stump rather than a tree. Richness of imagery is measured in all three activities.

Colorfulness of imagery. Creative products are usually those that are exciting and surprising. Colorfulness of imagery is the ability to create a drawing that induces sensations and emotions (Torrance, 1979). A drawing that includes some sort of unreal, spooky, earthy, magical, or a similar component receives a point for this creative strength (Torrance, 2000). Colorfulness of imagery is measured in all three activities.

Fantasy. Fantasies are rich resources that stimulate creative thinking (Torrance et al., 1992). According to Torrance et al. (1992), many creative individuals use fantasy when being involved in creative processes. Fantasy is the ability to include characters or objects from movies, books, fairy tales, myths, or a similar source or original fantasy in a drawing or title (Torrance, 2000). Giving human-like characteristics to objects (e.g., a talking tree) is another indication of this creative strength (Torrance et al., 1992). Fantasy is measured in all three activities.

Total Score on the Torrance Tests of Creative Thinking, Figural Forms

The TTCT-figural provide two types of test scores: the Average Standard Score and the Creativity Index. The Average Standard Score is estimated using the five norm-referenced variables, which are fluency, originality, elaboration, abstractness of titles, and resistance to premature closure. The Creativity Index, on the other hand, is estimated using all 18 variables measured in the TTCT-figural. Torrance et al. (1992) explained how to derive these test scores as summarized below.

Obtaining the Average Standard Score is a three-step process. First, raw scores for fluency, originality, elaboration, abstractness of titles, and resistance to premature closure are computed. The raw score of a variable is the sum of the points to which the variable has been given across the test. Raw scores include bonus points that are awarded for combinations of items. Second, both the national percentile and the standard score that correspond to a raw score are found for each of these five variables using the national percentile tables and the standard score tables by age or by grade. Finally, the standard scores of all five variables are added, and the Average Standard Score is obtained by dividing the total number by five. The Average Standard Score can then be converted to a percentile using the tables published by the Scholastic Testing Service.

Obtaining the Creativity Index is also a three-step process. First, each of the 13 creative strengths is given a score at the test level based on the number of expressions of the creative strength throughout the test. Second, scores that are given to all 13 creative strengths are added. Finally, the Average Standard Score and the overall creative strengths score are totaled. The sum of these two scores is assigned to an examinee as his or her Creativity Index. The Creativity Index can be converted to a percentile score from the published age or grade tables.

Standardization of the Torrance Tests of Creative Thinking, Figural Forms

The Scholastic Testing Service standardizes the TTCT-figural in the U.S. The most recent norms-technical manual of the TTCT-figural was published in 2008 (Torrance, 2008). The normative sample consisted of 70,093 examinees from kindergarten to college and above. The data were collected in 35 states. Of those 70,093 examinees, 7.4% were in the central U.S., 16.2% were in the northeastern U.S., 29.9% were in the southeastern U.S., and 46.5% were in the western U.S. No information on gender, ethnicity, or socioeconomic status about the sample was provided in the manual. The majority of the data were collected using Form A; the sample contained 54,788 examinees. Of those examinees, the majority of them were in kindergarten to fifth grade, totaling of 48,458 examinees.

Outside of the U.S., the standardization of the TTCT-figural is usually done by Torrance Centers™. In Turkey, there is currently no Torrance Center™; therefore, no national norms are available at this moment. A few scholars conducted studies on the TTCT-figural (e.g., Aslan, 2001; Aslan & Puccio, 2006), but these studies did not have large sample sizes that were enough to standardize the tests and publish the national norms in Turkey.

Reliability of Scores on the Torrance Tests of Creative Thinking, Figural Forms

Reliability is an essential requirement that an assessment tool must provide for its scores. Reliability assures that a test is able to deliver consistent and precise results whenever the test is administered under similar circumstances (AERA, APA, & NMCE, 2014; Crocker & Algina, 2008). The concept of reliability is described differently in different measurement theories. For instance, reliability coefficients and standard errors of measurement are used to explain reliability under the CTT framework. Under Rasch measurement theory, on the other hand, item information and test information functions are used for reliability.

Reliability studies on the TTCT-figural utilized the CTT framework. Therefore, these studies reported reliability coefficients to provide evidence for the reliability of scores on the tests. A considerable amount of research was conducted to examine the reliability of scores on the TTCT-figural. Interrater reliability, internal consistency reliability, alternate-forms reliability, and test-rest reliability were examined. Next, these studies will be summarized.

Starting with interrater reliability, the reliability coefficients reported in the test manual were at or above .95 for scores on each individual variable, the Average Standard Score, and the Creativity Index (Torrance, 2008). Several other studies also reported high interrater reliability coefficients ($r > .90$) for the Average Standard Score, the Creativity Index, and scores on each individual variable measured in the TTCT-figural (e.g., Ferracuti, Cannoni, Burla, & Lazzari, 1999; Johnson, 1985; Rudowicz, Lok, & Kitto, 1995; Torrance, 1972, 2000).

Regarding internal consistency reliability, evidence for three types of reliability (Kuder-Richardson 21, split-half, and coefficient-alpha) was reported in different studies. The most recent test manual (Torrance, 2008) reported Kuder-Richardson 21 reliability coefficients for internal consistency reliability. The reliability coefficients for the Average Standard Score and the Creativity Index were around .90 in the manual: The coefficients ranged from .85 to .94, with the coefficients for second grade being the lowest. Rampaul, Singh, and Didyk (1984) also reported high Kuder-Richardson 21 reliability coefficients ($r > .90$) for the total test score.

The most recent test manual (Torrance, 2008) did not report any other internal consistency reliability coefficients, but several studies examined split-half reliability and coefficient-alpha reliability. Prieto et al. (2006) estimated the split-half reliability coefficient for total scores based on the Parallel Lines task to be .92. Rampaul et al. (1984) reported that the split-half reliability coefficient for the total test score was around .90.

Split-half reliability coefficients for scores on the individual variables measured in the TTCT-figural were also estimated using the Spearman-Brown correction. Although the Spearman-Brown correction was used, split-half reliability coefficients for scores on the variables were mostly lower than .90 (Aslan & Puccio, 2006; Prieto et al., 2006; Rampaul et al., 1984). The coefficients ranged from .51 to .92, with the reliability coefficient for scores on resistance to premature closure reported by Aslan and Puccio (2006) being the lowest.

It should be noted that a considerably low number of items were used when the split-half reliability coefficients were estimated for some variables such as abstractness of titles and resistance to premature closure. For instance, regarding resistance to premature closure, there are 10 items in Activity 2, but the majority of the students cannot complete all 10 items within the 10-minute time limit. A potential scoring error made at the item level may also have degraded the reliability coefficients to a certain extent.

As cited in Krumm et al. (2014), Krumm and Lemos (2011) reported that the coefficient-alpha reliability coefficient for total scores on the TTCT-figural was .70, while Ferrando et al. (2007), Ferrando (2004), López (2001), and Prieto, López, Ferrándiz, and Bermejo (2003) reported that the coefficient-alpha reliability coefficient for total scores on the TTCT-figural was around .90. As similar with split-half reliability, the coefficient-alpha reliability coefficients for scores on the individual variables measured in the TTCT-figural were lower than .90 (Aslan & Puccio, 2006; Clapham, 2004): The coefficients ranged from .17 to .76, with the reliability coefficient for scores on resistance to premature closure reported by Aslan and Puccio (2006), again, being the lowest. It should be noted that a potential scoring error made at the item level (in addition to a low number of items included in the analyses) regarding some variables may have lowered the reliability coefficients for those variables.

The most recent test manual (Torrance, 2008) did not provide any evidence for alternate-forms reliability or test-retest reliability. However, several studies as well as the first two test manuals reported that alternate-forms reliability coefficients were around .80 (e.g., Dalbec, 1966; Hagender, 1967; Torrance, 1966, 1974) and that test-retest reliability coefficients were around .60 (e.g., Cropley & Clapson, 1971; Grover, 1963; Mackler, 1962). It should be noted here that these alternate-forms reliability and test-retest reliability coefficients were estimated for scores on fluency, originality, and elaboration as well as for total scores based on the variables measured from 1966 to 1984. Evidence for alternate-forms reliability and test-retest reliability is not available for the Average Standard Score, the Creativity Index, and scores on all the variables measured in the streamlined version of the TTCT-figural—the streamlined version has been in use since 1984.

Validity of Scores on the Torrance Tests of Creative Thinking, Figural Forms

Validity is another essential requirement that an assessment tool must provide for its scores. Validity addresses whether or not an assessment tool is capable of measuring what it is supposed to measure (AERA, APA, & NMCE, 2014; Cohen & Swerdlik, 1999). Although there are different types of validity, the validation process is considered to be cumulative (Hattie, Jaeger, & Bond, 1999). Each type of validity offers evidence for one type of inference of test scores and contributes to the validation process (AERA, APA, & NMCE, 2014; Crocker & Algina, 2008).

A number of studies examined the validity of scores on the TTCT-figural. Evidence for content validity, criterion-related validity, and construct validity will be summarized below due to their importance in the validation process. Studies of predictive validity will be summarized to

provide evidence for criterion-related validity. Studies of factor analysis and discriminant validity will be summarized to bring evidence for construct validity.

Starting with content validity, Torrance (1979) and Torrance and Safter (1999) made strong arguments about the content of the TTCT-figural and provided explanatory evidence for content validity. In addition, numerous articles and books that discussed the assessment of creativity mentioned the TTCT-figural as assessment tools that measure creative thinking skills (e.g., Hunsaker, Abeel, & Callahan, 1991; Kaufman, Plucker, & Russell, 2012; Plucker & Makel, 2010; Sternberg, 2006). These support that the content of the TTCT-figural cover an adequate portion of creative thinking skills and that the tests are recognized as instruments that measure creative thinking skills.

Regarding predictive validity, it was reported in the most recent test manual that the test scores were significantly correlated with adult creative achievements (r = .51; Torrance, 2008). Torrance (2000) reported similar results: Predictive validity coefficients were around .50 for scores on the variables measured in the streamlined version of the TTCT-figural. The predictive validity coefficients estimated 40 years and 50 years after the first testing were also satisfactory: The index of creativity which was based on fluency, originality, elaboration, and flexibility—flexibility was measured from 1966 to 1984—was moderately correlated with creative achievements 40 years (r = .43; Cramond, Matthews-Morgan, Bandalos, & Zuo, 2005) and 50 years after the first testing (r = .40; Runco, Millar, Acar, & Cramond, 2010).

Evidence for the factorial structure of the TTCT-figural is inconsistent. One-factor, two-factor, three-factor, and four-factor models were identified in different studies (e.g., Aliotti & Blanton, 1969; Antunes & Almeida, 2007; Auzmendi et al., 1996; Clapham, 1998, 2004; Ferrando, 2006; Heausler & Thompson, 1988; Kim, 2006; Kim, Cramond, & Bandalos, 2006;

Krumm et al., 2014; Oliveira, 2007; Prieto et al., 2006). It is worth noting here that in these factor analysis studies, analyses were conducted at the variable level, not at the item level for each individual variable measured in the TTCT-figural.

Another important point regarding the findings of factor analysis studies is that studies that identified the same numbers of factors did not identify the same factorial structures (e.g., Heausler & Thompson, 1988; Kim, 2006; Krumm et al., 2014). For instance, both Heausler and Thompson (1988) and Krumm et al. (2014) identified a two-factor model; however, Heausler and Thompson found that one factor loaded on fluency ($\lambda = .63$), originality ($\lambda = .84$), elaboration ($\lambda = .72$), and resistance to premature closure ($\lambda = .70$) and that another factor loaded on abstractness of titles ($\lambda = .83$); whereas Krumm et al. found that one factor loaded on fluency ($\lambda = .63$) and originality ($\lambda = .96$) and that another factor loaded on elaboration ($\lambda = .26$), resistance to premature closure ($\lambda = .84$), and abstractness of titles ($\lambda = .58$).

Finally, studies that examined discriminant validity found weak correlations ($r < .20$) between scores on the TTCT-figural and scores on intelligence tests (e.g., Cho, Nijenhuis, van Vianen, Kim, & Lee, 2010; Crawford & Nirmal, 1976; Esquivel & Lopez, 1988; Palaniappan, 2008; Yong, 1994). Considering the notion that the TTCT-figural primarily require divergent thinking (i.e., thinking in different directions and generating several ideas) whereas intelligence tests require convergent thinking (i.e., thinking in one direction and finding the right idea), the findings of these studies provided evidence that the TTCT-figural measure different constructs from intelligence tests, as expected.

Fairness of Scores on the Torrance Tests of Creative Thinking, Figural Forms

Whenever an assessment tool is used to assess the trait levels of examinees of different cultures, socioeconomic status, ethnicities, and genders, it is essential to show that the

assessment tool does not discriminate against any particular group of examinees. For this particular purpose, several studies examined the fairness of scores on the TTCT-figural in terms of gender (e.g., Awamleh, Al Farah, & El-Zraigat, 2012; Campos, Lopez, Gonzalez, & Perez-Fabello, 2000; Cheng, Kim, & Hull, 2010; Kim et al., 2006; Kim & VanTassel-Baska, 2010; Matud & Grande, 2007; Rudowicz et al., 1995), ethnicity (e.g., Cheng et al., 2010; Palaniappan, 2008; Saeki, Fan, & Van Dusen, 2001; Sikka, 1992; Tannehill, 1992; Tran, 2004), and socioeconomic status (e.g., Johnson, 1974; Ogletree & Ujlaki, 1973; Voss, 1997).

On no occasion did examinees in one group (e.g., females or high socioeconomic status) consistently score significantly higher on the TTCT-figural than did examinees in that group's counterpart (e.g., males or low socioeconomic status). In some study settings, examinees in one group scored higher than examinees in its counterpart; however, in other study settings, the difference shifted or disappeared (e.g., Cheng et al., 2010; Kim & VanTassel-Baska, 2010; Matud & Grande, 2007; Rudowicz et al., 1995). For instance, Matud and Grande (2007) found that males scored significantly higher than females on originality, whereas Cheng et al. (2010) found that females' originality scores were significantly higher than those of males. Findings of studies on the fairness of scores on the TTCT-figural suggested that the TTCT-figural did not favor any particular group of examinees.

Interim Conclusion on the Torrance Tests of Creative Thinking, Figural Forms

Studies that examined the reliability, validity, and fairness of scores on the TTCT-figural showed that the tests deliver reliable Average Standard Scores and Creativity Index Scores; that evidence for content validity, predictive validity, and discriminant validity is satisfactory; and that the tests do not discriminate against any particular group of examinees in terms of gender, ethnicity, and socioeconomic status. Low internal consistency reliability coefficients ($r < .60$) for

particular variables, such as abstractness of titles and resistance to premature closure, indicated that the internal consistency reliability of scores on some variables might lack sufficient reliability and that each variable needs to be studied in detail so that the standard errors of measurement involved in scores on each variable can be estimated.

It should be noted, however, that the reliability coefficients for several variables (e.g., abstractness of titles and resistance to premature closure) were affected severely by the number of items included in the analyses. For instance, abstractness of titles is scored on 11 items and resistance to premature closure is scored on 10 items, but not all students can produce 11 responses for abstractness of titles and 10 responses for resistance to premature closure. A potential scoring error made at the item level may also have lowered the reliability coefficients for some variables.

The existing evidence for alternate-forms reliability and test-retest reliability is acceptable, but the absence of data on alternate-forms reliability and test-retest reliability of scores on the streamlined version of the TTCT-figural indicated that further research on these two types of reliability is needed. Finally, evidence for the factorial structure of the TTCT-figural is not satisfactory. Inconsistent factorial structures identified in factor analysis studies suggested that the factorial structure of the tests needs to be examined at the item level for each individual variable.

Studies of validity, reliability, and fairness provided no evidence to support the seven assumptions about the TTCT-figural and to confirm that the TTCT-figural deliver invariant calibration of the items as well as invariant measurement of the creative thinking skill levels of examinees. In conclusion, studies on the TTCT-figural showed that CTT and factor analysis cannot address all seven assumptions about the tests and that another measurement theory,

preferably Rasch measurement theory, must be utilized to test the seven assumptions about the TTCT-figural and to examine whether the tests deliver invariant calibration of the items as well as invariant measurement of the creative thinking skill levels of examinees.

Applications of Rasch Measurement Theory in Creativity Testing

Although Rasch measurement theory offers major advantages for both testing and test development, a few researchers have applied Rasch measurement theory to creativity testing and the development of creativity-related instruments (see Ariffin, Katran, Badib, & Rashid, 2011; Karwowski, 2014; Nakano & Primi, 2014; Wang, Ho, Cheng, & Cheng, 2014; Wechsler, Vendramini, & Oakland, 2012). The majority of these studies examined questionnaires and scales that measure creativity-related skills or creative achievements. Of these studies, only Nakano and Primi (2014) used Rasch measurement theory to examine a test that has the same structure as the TTCT-figural. All these studies will be summarized below.

Studies on Creativity-Related Instruments

Ariffin, Katran, Badib, and Rashid (2011) used the dichotomous Rasch model and rating scale model to examine the Malaysian Creativity and Innovation Instrument (MyCrIn; Ariffin et al., 2011). MyCrIn contains 290 items and is comprised of five subtests each of which measures a different construct in a self-report format. These five constructs are higher-order thinking, curiosity, sensitivity, being visionary, and being adaptable to change. Some of the items in the instrument are dichotomous items, and some of the items are polytomous items. The participants of the study were 285 Malaysian university students. Item analyses were conducted to examine item fit, item difficulty, average item difficulty, and differential item functioning.

Analyses of item fit showed that 11 items had misfit (i.e., underfit), indicating that at least one other factor other than the trait measured by each item was affecting the examinee responses.

Item difficulty levels for the dichotomously scored items were computed to range from -3.50 to 3.50 logits; however, the majority of the items were clustered between -2.00 and 2.00 logit points. Average item difficulty for the polytomously scored items ranged from -1.50 to 2.50, but the majority of the items were between -1.00 and 1.00 logits in average difficulty. Difficulty levels of the response options for the polytomously scored items were not reported in the study. Differential item functioning was investigated for gender. Analyses showed that eight items behaved differently for each gender: Six of those items were in favor of females, while two of them were in favor of males.

Wang, Ho, Cheng, and Cheng (2014) examined the Creative Achievement Questionnaire (CAQ; Carson et al., 2005) using the dichotomous Rasch model. CAQ is a self-report instrument which measures creative achievements in 10 different domains (visual arts, music, dance, creative writing, architectural design, humor, invention, scientific inquiry, theater and film, culinary arts). Each domain consists of eight yes/no questions. The developers of the instrument hypothesized that the questions in each domain gradually became more difficult from the first question to the last with respect to the level of achievement they measured. In the study, the questionnaire was administered to 905 Taiwanese participants whose ages ranged from 14 to 78. Item analyses were conducted to examine item fit, item difficulty, and differential item functioning.

Item fit measures showed that all the items had a good fit, indicating that the requirements for unidimensionality and local independence were met. Overall, the level of item difficulty ranged from -5.46 to 3.57 logits for all the items in the questionnaire. For the majority of the domains, item difficulty level gradually increased within a domain from the first question to the last, as suggested by the developers. However, in some domains (e.g., dance and creative

writing), item difficulty levels had hierarchical fluctuations (i.e., item difficulties did not increase gradually within a domain from the first question to the last). This finding implied that the items in these domains need to be revised. Finally, analyses of differential item functioning were conducted for gender. Analyses revealed that nine of the questions in the questionnaire had different difficulty levels for males and females, indicating that these items did not behave the same for each gender.

Wechsler, Vendramini, and Oakland (2012) used the partial credit model to examine the Styles of Thinking and Creating (STC; Wechsler, 2006). The STC is a self-report measure which contains 100 6-point Likert-type items and measures five thinking styles: cautious-reflexive, nonconforming-transformator, logical-objective, emotional-intuitive, and relational-divergent. The response options are: totally disagree, disagree, partially disagree, partially agree, agree, and totally agree. The instrument was given to 1,752 Brazilian participants whose ages ranged from 17 to 70. Item analyses were conducted to examine item fit and the difficulty levels of the response options (i.e., item scaling).

Twenty-seven items on the instrument showed misfit (i.e., underfit). This finding indicated that unidimensionality might be violated for those misfitting items and that a second factor might have impacted the student responses given to those items. Analyses revealed that the difficulty level of the third response option, partially disagree, was not significantly different from the second response option, disagree. Similarly, the difficulty level of the fourth response option, partially agree, was not significantly different from the fifth response option, agree. Therefore, the researchers concluded that the third and fourth response options could be removed and that a 4-point scale could be used for all the items in the instrument.

Karwowski (2014) examined the Creative Mindset Scale (CMS; Karwowski, 2014) using the rating scale model. The CMS is a self-report scale which consists of 10 5-point Likert-type items and measures perception of creativity. Five of the items measure growth-mindset of creativity, while the other five items measure fixed-mindset of creativity. The scale was given to 699 participants whose ages ranged from 16 to 60. Item analyses were conducted to examine test information.

Because the scale contains Likert-type items, the test information had a wide peak. The test information peaked between -2.00 and 1.00 logit points, indicating that the total scores around the average score were the most reliable. The test information decreased towards the high and low ends of the latent trait continuum but the decrease was larger at the high end, indicating that the high total scores were the least reliable and that the low total scores were in between the average scores and high scores in terms of reliability.

In addition to the test information, the total amount of information provided by the items on growth-mindset of creativity and fixed-mindset of creativity was estimated. The total amount of information provided by the five items measuring the growth-mindset of creativity was the highest around the -1.50 logit point. The total amount of information provided by the other five items measuring the fixed-mindset of creativity was the highest around the 0 logit point.

Study on the TTCT-Figural-Like Instrument

Nakano and Primi (2014) worked with 1,426 students from ages 6 to 15 to examine a test called the Test of Creativity in Children's Drawings (TCCD; Nakano, Wechsler, & Primi, 2011). The TCCD is a figural test of creative thinking based on the TTCT-figural Form A. The partial credit model was used for analyses. Analyses were conducted to examine item fit and the difficulty levels of the response options for each item (i.e., variable).

The TCCD measures 12 creative thinking skills in three activities. Like the TTCT-figural Form A, Activity 1 in the TCCD has one item, Activity 2 has 10 items, and Activity 3 has 30 items. The creative thinking skills (i.e., variables) measured in the test are fluency, flexibility, elaboration, originality, emotional expression, fantasy, movement, unusual perspective, internal perspective, use of context, extension of limits, and expressive titles. Scores on the TCCD are highly correlated with scores on the TTCT-figural (r = .91; Nakano & Primi, 2014).

Unconventionally, Nakano and Primi (2014) did not examine the individual items on the test regarding each variable. Instead, the researchers used total scores on each individual variable as item scores and conducted analyses for each variable as if it were an item. This technique is referred to as pseudoscaling, and it might offer a good understanding of the nature of the variables measured in a test and allow researchers to make inferences about the variables (Bond & Fox, 2001).

In order to use a particular variable as an item, Nakano and Primi (2014) created a Likert scale for the variable. The number of the response options of this Likert scale was determined based on the maximum possible points that can be attained on the variable. For instance, elaboration (i.e., elaboration 3) was a 10-point Likert-type item in Activity 3, while elaboration (i.e., elaboration 2) was a 5-point Likert-type item in Activity 2.

Analyses showed that the majority of the items had a good fit. Both the infit and outfit mean square values for three items, including fluency 3, elaboration 1, and expressive titles 2, were below 0.80, the cutoff value recommended for a good fit (Engelhard, 2013), meaning that these items showed overfit. When Rasch analyses are conducted at the item level, good item fit indicates that responses given to the items on a test are affected by only one factor and are independent of each other (Bond & Fox, 2001). Overfit, on the other hand, indicates that

responses are too deterministic and that local independence might be violated (Bond & Fox, 2001).

Because Nakano and Primi (2014) conducted analyses at the variable level, not at the item level, straightforward interpretations are difficult to make. One possible interpretation regarding overfitting items is that if one is to create a pseudoscale for each variable, all the variables, except fluency 3, elaboration 1, and expressive titles 2, provide appropriate measurements of the creative thinking skill levels of examinees. Another possible interpretation is that if one is to create a pseudoscale for each variable, fluency scores in Activity 3, elaboration scores in Activity 1, and expressive titles scores in Activity 2 can be predicted by scores on the same variables in the other two activities.

Regarding item difficulty, analyses revealed that the response options were not distributed evenly on the latent trait continuum for the majority of the items. For instance, the fourth, fifth, sixth, seventh, eighth, and ninth response options for elaboration 3 were clustered within 1 logit, while the first, second, and third response options were also clustered within 1 logit. These findings indicated that for some items, the response options were not based on equal intervals, as assumed by the developers and that some response options might need to be collapsed after further research.

In the study, Nakano and Primi (2014) did not examine the actual items on the test. However, the actual test items should have been analyzed for each individual variable before conducting variable-level analyses. Without analyzing the actual items on the test, the researchers essentially assumed that each individual item on the test was equally difficult and that item fit was good for all the items, although these assumptions might in fact not be true.

The findings of the study, however, are still promising. The study demonstrated that Rasch measurement theory could offer a good understanding of the quality of the variables measured in the TTCT-figural and that Rasch measurement theory analyses could examine the assumption about elaboration. Although Nakano and Primi's (2014) study did not directly demonstrate that Rasch measurement theory could be applied to the TTCT-figural to examine the individual items for each variable, there is no reason to avoid conducting analyses on the actual items in the TTCT-figural to examine the tests.

CHAPTER 3

METHOD

Sample

The participants of the study were second grade, age 7, students in Turkey. The students came from four different public schools each of which is located in a different school district in Istanbul or Tekirdag. Overall, 193 students, 109 girls and 84 boys, participated in the study: Twenty-six of them were in the Bakirkoy school district in Istanbul, 48 of them were in the Esenyurt school district in Istanbul, 61 of them were in the Esenler school district in Istanbul, and 58 of them were in the Suleymanpasa school district in Tekirdag.

These four school districts were chosen due to the socioeconomic status (SES) profiles of parents in these districts. The Bakirkoy school district is a high SES area, the Esenler and Suleymanpasa school districts are average SES areas, and the Esenyurt school district is a low SES area. By conducting the study in these school districts, students from high, average, and low SES families were included. To which SES group a school district belonged was determined based on the cost of housing in that particular area.

Instrument

The TTCT-figural Form A (Torrance, 1984) was administered in the study. The instrument contains three activities, each containing a different number of items. In all three activities, examinees are asked to create a figural response using each individual item in the activity as the starting point and to give a title to each response (Torrance, 2006a, 2006b). Examinees are given 10 minutes to respond in each activity.

Data Collection

Before data collection, permission from the local bureaus of the Ministry of National Education in Istanbul and Tekirdag was obtained so that different schools could be approached. I initially contacted three public schools in Istanbul and one public school in Tekirdag, which all eventually allowed me to conduct the study. After receiving permission from these schools, the parents of the students were notified that their child might participate in a study on creativity.

I personally visited the schools and administered the tests. Before the tests were administered, the students' creative thinking skills were stimulated with a warm-up activity. In this warm up activity, the students were asked "In how many different ways can you use a shoe?" Afterwards, the students were given the TTCT-figural Form A. The tests were administered in groups, each containing 14 to 18 students—classrooms with students over 20 were divided into two groups. The students worked on Activity 1 and Activity 2 subsequently, and then they were given an 8-minute break. After this break, they worked on Activity 3.

The 10-minute time limit given for each activity was enough for the students to create a response in Activity 1 and a sufficient number of responses in Activity 3. Thus, the students were given 10 minutes for these two activities. However, the majority of the students could not finish all 10 items in Activity 2 within the 10-minute time limit, as expected. Therefore, I gave those students who did not finish all 10 items in Activity 2 up to 15 minutes of additional time to obtain a sufficient number of responses for each item.

The students used a different color of pen to complete the activity during the supplementary time given for Activity 2, which was granted after the students worked on Activity 3. This decision (i.e., giving a supplementary time for Activity 2) was based on Cohen and Swerdlik's (1999) suggestion regarding analyzing the items on a timed test. Cohen and

Swerdlik recommended that researchers provide additional time to all examinees to complete all the items on a timed test if the response rate is not the primary interest.

Scoring Process

Because I was trained on how to score the TTCT-figural, I scored the tests. The tests were scored based on the TTCT-figural scoring manual (Torrance et al., 1992). Meaningful responses received the predetermined point(s) explained in the second chapter. Meaningless responses were not treated as missing and received 0 points.

Burak Turkman, a certified rater, scored 10% of the tests to check the reliability of scoring. Burak Turkman's item scores and my item scores were correlated to estimate the interrater reliability coefficients for the variables—the coefficient for elaboration was based on the activity scores. The interrater reliability coefficients were as follows: .99 for fluency, .93 for originality, .91 for elaboration, .84 for elaboration-I, .89 for abstractness of titles, .89 for resistance to premature closure, .89 for emotional expressiveness, .90 for storytelling articulateness, .89 for movement or action, .90 for expressiveness of titles, 1.00 for synthesis of incomplete figures, 1.00 for synthesis of lines, .87 for unusual visualization, .90 for internal visualization, .91 for extending or breaking boundaries, .85 for humor, .87 for richness of imagery, .86 for colorfulness of imagery, and .91 for fantasy.

Because elaboration-I was fabricated for analyses purposes in the current study, there is no information in the test manual on how to score elaboration-I for each individual item in Activity 2 and Activity 3. When the test items were scored on elaboration-I, the following criteria were used: For up to five details, the number of details in an item was given to the item as its elaboration-I score. For instance, if there were zero details, the item was given 0 points; if there was one detail, the item was given 1 point. If there were six or more details in a response

given to an item, the item was given 6 points. This type of scoring was employed instead of counting the actual details added to a response given to an item and using the number of details as an item score because the model used to analyze the items regarding elaboration-I requires each test item to have the same number of response options.

There were several missing observations in the study. A number of students did not create figural responses using some of the items and left them untouched or forgot to give a title to a figural response. Those unanswered items and absent titles were treated as missing. The missing observations were random. Because the software used for analyses can run the analyses when there are missing data, all the missing data were included in the analyses as missing.

Analyses

When the activity scores were calculated for elaboration prior to conducting analyses, only the items completed within the 30-minute time limit were used to estimate the elaboration score of an examinee for each activity. For instance, if a student responded to the first five items in Activity 2 and first eight items in Activity 3, only those items were used to estimate the elaboration score for each activity. This type of scoring was done for only elaboration because elaboration is the only variable scored at the activity level, and it is the ability to add details within the 10-minute time limit given for each activity (Ball & Torrance, 1984). If the items completed during the supplementary time had been included, the number of details in Activity 2 and Activity 3 would have been inflated, so the lower response options would have been constrained within a small trait range and thus become meaningless. Also, if the items completed during the supplementary time had been included, Activity 2 and Activity 3 would not have been analyzed under the same conditions as Activity 1 for elaboration.

For the rest of the variables, the item in Activity 1, all 10 items in Activity 2, and the first six items in Activity 3 were analyzed. Only the first six items in Activity 3 were analyzed for those variables because although all the items in Activity 3 are pairs of parallel lines, it was possible that the location of a pair of parallel lines and the distance between the lines might impact how the pair was perceived and how it was used to create a response. Thus, I analyzed the first six items in Activity 3 for all the variables measured in the activity. Ideally, all 30 items in Activity 3 should have been analyzed; however, I was concerned that the students—because they were seven years old—might get tired and lose their attention while working on all 30 items or run out of ideas. This might have impacted the validity of data collection process.

Item analyses were conducted using models described in Rasch measurement theory. The dichotomous Rasch model (Rasch, 1960) and the rating scale model (RSM; Andrich, 1978) were used to examine the research questions. The dichotomous Rasch model was used to analyze the items regarding the dichotomously scored variables. The RSM was used to analyze the items and activities regarding the polytomously scored variables. Data analyses were conducted on Facets software (Linacre, 2013).

The Dichotomous Rasch Model

The dichotomous Rasch model is based on Rasch's (1960) statistical computations. The model is appropriate to use when the items on a test are scored in a dichotomous manner such as original vs. unoriginal. Under the dichotomous Rasch model, the items on a test must have the same level of item discrimination but they can differ from each other in the level of item difficulty (Bond & Fox, 2001; Engelhard, 2013). Item parameters as well as person parameters are independent of each other under the dichotomous Rasch model (Engelhard, 2013). Therefore,

the dichotomous Rasch model allows researchers to study invariant calibration of the items on a test as well as invariant measurement of the trait levels of examinees.

It is hypothesized under the dichotomous Rasch model that there is a probabilistic relation between the difficulty level of an item and the trait level of an examinee (Embretson & Reise, 2000; Engelhard, 2013). Accordingly, the difference between the examinee's trait level and the item's difficulty yields the probability of the examinee being successful on the item (Engelhard, 2013; Bond & Fox, 2001). The dichotomous Rasch model utilizes the following equation to estimate the probability of success on a particular item for an examinee:

$$P_{ni}(x = 1|\theta_n, \delta_i) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad (1)$$

In this equation, *P* is the probability of examinee *n* correctly answering (i.e., *x*=1) item *i*, θ is the trait level of examinee *n*, δ is the level of difficulty for item *i*, and exp is the exponential function which raises the value of e (≈2.718) to the power of the difference between the examinee *n*'s trait level (θ_n) and the item *i*'s difficulty (δ_i).

The Rating Scale Model

The RSM was developed by Andrich (1978). The model is an extension of the dichotomous Rasch model; therefore, item discriminations for all the items on a test must be equal (Embretson & Reise, 2000), and item parameters as well as person parameters are independent of each other (Engelhard, 2013). The RSM is appropriate to use when the items on a test are comprised of ordered response options and the response options have the same rating scale structure across the items (Bond & Fox, 2001; Engelhard, 2013).

Because the items on a test have the same rating scale structure, a single item location parameter (i.e., average item difficulty) can be estimated for each individual item on the test in addition to a difficulty parameter for each threshold (Bond & Fox, 2001; Embretson & Reise,

2000). Thresholds are points on the latent trait continuum where two subsequent response options have the same probability of being responded to by examinees (Bond & Fox, 2001).

RSM analyses allow researchers to identify the locations of the response options of a polytomously scored item on the latent trait continuum. Therefore, the RSM yields true interval scales (Bond & Fox, 2001; Engelhard, 2013). In addition, using the RSM, one can examine whether the polytomously scored items on a test have the same distribution structure and whether the response options of the items are based on equal intervals (Engelhard, 2013). Engelhard (2013) described the rationale behind using the RSM as follows:

> The basic idea that motivates the use of Rasch models for rating scale data is that the scoring of categories with ordered integers (0, 1, …, *m*) may be based on the unexamined assumption that the categories define equal intervals on the latent variable. This assumption may not be justified when empirical information is examined. Rasch models provide a framework to explicitly examine this assumption and to parameterize the intervals that define the categories without the assumption that the categories are of equal size. (p. 49)

Similar to the dichotomous Rasch model, it is hypothesized under the RSM that there is a probabilistic relation between the trait level of an examinee and the difficulty level of an item as well as the difficulty level of a particular threshold (Bond & Fox, 2001; Engelhard, 2013). The RSM utilizes the following equation to estimate the probability of success on a response option:

$$P_{nik}(x = 1|\theta_n, \delta_i, \tau_k) = \frac{\exp(\theta_n - \delta_i - \tau_k)}{1 + \exp(\theta_n - \delta_i - \tau_k)} \quad (2)$$

In this equation, *P* is the probability of examinee *n* responding to response option *k* (i.e., *x*=1) in item *i*, θ is the trait level of examinee *n*, δ is the level of difficulty for item *i*, τ is the difficulty level of the *k*th threshold, and exp is the exponential function. The difficulty level of

the *k*th threshold indicates the minimum required trait level to respond to response option *k* (Bond & Fox, 2001).

Item Parameters

Rasch measurement theory offers several item parameters. These parameters are item fit, item difficulty or average item difficulty, Rasch-Andrich thresholds, item information, and item reliability index. Under Rasch measurement theory, item parameters are the group of parameters that are estimated first. Facets software ignores the person parameters and estimates the item parameters first (Bond & Fox, 2001). Based on the responses given to an item, Facets determine the location (i.e., difficulty level) of the item on the latent trait continuum and calculates the other item parameters. Item parameters will be explained below.

Item fit. In Rasch measurement theory, item fit is an indication of the quality of an item (Bond & Fox, 2001; Embretson & Reise, 2000). Researchers use item fit to evaluate whether an item has any issues that could cause inadequate measurements of the trait levels of examinees. Fit statistics show whether a particular item on a test has a good fit or misfit. Good item fit implies that the responses given to an item are affected by one factor and independent of each other (Bond & Fox, 2001). Misfit, on the other hand, indicates that either unidimensionality or local independence might be violated (Bond & Fox, 2001). Item fit statistics are summarized using the mean squares method and the *t* distribution method (Engelhard, 2013).

The mean squares method involves two sub methods: the infit mean square method and the outfit mean square method. In both methods, the value of 1.00 is the projected mean square value, and it corresponds to the exact amount of variation predicted by Rasch measurement theory (Wright & Linacre, 1994). Any infit or outfit mean square value above 1.00 indicates more variation in the item measures than Rasch measurement theory predicted, while an infit or

outfit mean square value below 1.00 implies less variation than Rasch measurement theory projected (Wright & Linacre, 1994). For instance, an infit or outfit mean square value of 1.15 means that there is 15% more variation in the item measures than Rasch measurement theory predicted.

An item is considered to have a good fit if the infit or outfit mean square value is between 0.80 and 1.20 (Bond & Fox, 2001; Engelhard, 2013; Wilson, 2005; Wright & Linacre, 1994). Out of this range, the item is considered to have misfit. Good item fit denotes some degree of variation in the item measures, but this variation does not impact the assessment of the trait levels of examinees. An infit or outfit mean square value smaller than 0.80 indicates that the responses given to the item are too deterministic and that local independence might be violated (Bond & Fox, 2001). This type of misfit is referred to as overfit. An infit or outfit mean square fit value bigger than 1.20 implies that the responses given to the item are too haphazard and that unidimensionality might be violated (Bond & Fox, 2001). This type of misfit is called underfit.

Wright and Linacre (1994) opined that underfit was a much greater threat than overfit for the assessment of the trait levels of examinees because underfitting items produce unpredictable response patterns. On the other hand, regarding overfit, infit or outfit mean square values below 0.80 "do not contradict what we know [about the items], but they do not tell us much that is new about what we want to know" (Wright & Linacre, 1994, p. 370).

There are two sub methods of the *t* distribution method: the infit *t* distribution (i.e., standardized infit) method and the outfit *t* distribution (i.e., standardized outfit) method. Zero is the projected value for both the standardized infit and outfit methods, and it indicates that a good item fit is highly possible for the item (Bond & Fox, 2001). Any standardized infit or outfit value between -2.00 and 2.00 indicates that a good item fit is likely for the item (Bond & Fox, 2001).

Out of this range, misfit becomes probable. Below -2.00, overfit is more possible than a good fit. Above 2.00, on the other hand, underfit becomes more probable than a good fit.

In the present study, the infit mean square method and the standardized infit method were used to evaluate item fit. Infit statistics (i.e., infit mean square method and standardized infit method) were chosen because outfit statistics (i.e., outfit mean square method and standardized outfit method) are heavily influenced by the extreme responses in the data, whereas infit statistics are not vulnerable against the extreme responses in the data (Linacre, 2013). This is an important distinction regarding the present study because there were many students in the study with a perfect score or zero score on a particular variable. Infit statistics were more appropriate to use comparing to outfit statistics.

Analyses of item fit in the current study examined whether a particular variable was the only factor affecting the responses given to each item or each activity and whether the responses given to the items were too deterministic regarding a particular variable. In other words, analyses of item fit examined construct validity of the TTCT-figural Form A and local independence at the item level for the variables measured in the test.

Item difficulty. Item difficulty is estimated for the dichotomously scored items on a test, while average item difficulty is estimated for the polytomously scored items on a test (under the rating scale model). In general, item difficulty shows how difficult it is to correctly answer a dichotomously scored item (Bond & Fox, 2001). Similarly, average item difficulty indicates how difficult a polytomously scored item is with respect to the variable measured by the item (Bond & Fox, 2001).

Item difficulties are expressed in logits in Rasch measurement theory. Logits are units of measurement that are based on an interval scale. Difficulty parameters are constrained to a

particular range such as between -3 and +3, with a mean value of 0. Difficulty levels increase as the value of item difficulty goes from negative to positive.

With respect to the TTCT-figural, item difficulty indicates how challenging it is to manifest a particular dichotomously scored variable in an item. For instance, regarding originality, item difficulty shows how challenging it is to create an original response using a particular item. In a similar fashion, average item (or activity for elaboration) difficulty indicates how challenging an item (or activity) is with regard to a polytomously scored variable. For instance, regarding resistance to premature closure, average item difficulty shows how challenging each item in Activity 2 is.

In the current study, analyses of item difficulty examined whether it was equally difficult to manifest a particular dichotomously scored variable in each individual item. Analyses of item average difficulty, on the other hand, examined whether each test item was equally difficulty with respect to a polytomously scored variable. Regarding elaboration, analyses of average item (i.e., activity) difficulty tested whether each individual activity was equally difficult.

Category thresholds. Category thresholds are also known as Rasch-Andrich thresholds in Rasch measurement theory. Rasch-Andrich thresholds are points on the latent trait continuum where two subsequent response options have the same probability of being responded to by examinees (Bond & Fox, 2001). Rasch-Andrich thresholds show how difficult each response option of a polytomously scored item is. Below a threshold point, the lower response option is more probable to be responded to by an examinee; while above the threshold point, the higher response option is more probable to be responded to by the examinee.

A Rash-Andrich threshold can be thought of as a separation point which separates one response option from another response option. For instance, there are four possible response

options of the items regarding abstractness of titles. Therefore, there are three thresholds in each item regarding abstractness of titles. The first threshold separates the first and second response options, the second threshold separates the second and third response options, and the third threshold separates the third and fourth response options.

In addition to the Rasch-Andrich thresholds, Facets also estimates the lowest trait level and the highest trait level measured by an item. The lowest trait level and the highest trait level measured by an item indicate the operational range of the item. The lowest trait level measured by an item shows at what point on the latent trait continuum the item starts measuring the variable. Similarly, the highest trait level measured by an item shows at what point on the latent trait continuum the item stops measuring the variable. In the current study, Rasch-Andrich thresholds and operational ranges were estimated to examine whether the response options of the items (or activities regarding elaboration) were distributed equally on the latent trait continuum regarding a polytomously scored variable.

Item information. In Rasch measurement theory, item information is tied to reliability (Baker, 2001; Cappelleri et al., 2014). The amount of information an item provides on the measured variable is estimated through the item information function. The item information function allows researchers to estimate how much standard error of measurement is involved in an item at a particular trait level. As the amount of information an item provides on the measured variable increases, the amount of standard error of measurement involved in the item decreases.

The amount of information provided by individual items on a test at a particular trait level can be totaled to estimate the amount of information provided by the test at that trait level (Baker, 2001; Embretson & Reise, 2000). Because the test information function is used to estimate how much standard error of measurement is involved in the test scores at each point on

the latent trait continuum, Rasch measurement theory can offer different standard errors of measurement for different test scores.

Analyses of item information in the current study showed whether each individual item provided the same amount of information on a particular variable scored at the item level and whether the amount of information a particular item provided on this variable was at the same level along the latent trait continuum. For elaboration, analyses of item information examined whether each individual activity provided the same amount of information and whether the amount of information a particular activity provided on elaboration was at the same level along the latent trait continuum.

Item reliability index. Rasch measurement theory offers two types of reliability index: person reliability index and item reliability index (Bond & Fox, 2001). It should be noted that the term reliability here refers to invariant measurement which has a different meaning from the term *reliability* meaning the consistency of test scores. The item reliability index is used to assess whether an instrument delivers sample-independent (i.e., invariant) item calibration (Bond & Fox, 2001).

The item reliability index can have a value between 0 and 1.00. The cutoff point recommended for invariant item calibration is .90 (Linacre, 2016). Any value above .90 indicates a sufficiently high possibility that invariant item calibration is achieved.

The item reliability index implies whether the item parameters (e.g., item difficulty) will remain at the same level if the instrument is administered to another group of examinees (Wright & Masters, 1982). In the current study, the item reliability indexes showed whether the items in the TTCT-figural Form A could be calibrated in a sample-independent manner regarding a particular variable. Additionally, the item reliability indexes indicated whether the item

parameters would remain at the same level if the test was given to another group of second grade students.

Person Parameters

Rasch measurement theory offers a few person parameters. Specifically, person measures (i.e., trait levels of examinees), person fit, and person reliability index are estimated under the Rasch measurement theory framework. As previously mentioned, Facets software first estimates the item parameters (e.g., item difficulty) by ignoring the person parameters. After the item parameters are obtained, the person parameters are estimated (Bond & Fox, 2001). Person parameters will be explained below.

Person measures. Based on the responses given by an examinee to the items on a test, Facets determine the most probable location of the examinee on the latent trait continuum (i.e., the trait level of the examinee) by taking the difficulty level of each item on the test into consideration (Bond & Fox, 2001; Embretson & Reise, 2000). Like item difficulties, person measures are also expressed in logits. Therefore, both the difficulty levels of the items on a test and the trait levels of the examinees taking the test can be displayed on the same latent trait continuum. In Rasch measurement theory, researchers use variable maps for this purpose.

Although the location (i.e., trait level) of each student participating in the present study was determined, person measures will not be discussed in detail in the following sections because the primary objective of the study was to examine the quality of the items in the TTCT-figural Form A—person fit and person reliability index will be critiqued to evaluate the quality of the items in the TTCT-figural Form A as a group.

Person fit. Under the Rasch measurement theory framework, person fit has a similar purpose with item fit (Wright & Linacre, 1994). Person fit is an indication of the accuracy of the

person measures. Person fit statistics are also used to evaluate the quality of the items on a test as a group. Person fit is used to identify whether a set of items (i.e., test) yield too haphazard or too deterministic response patterns (Wright & Linacre, 1994).

Person fit statistics show whether a person measure (i.e., trait level of an examinee) has a good fit or misfit. Good person fit implies that the variation in the person measure is acceptable. Misfit for person measures, on the other hand, indicates that there is more or less variation in the person measures than Rasch measurement theory predicted.

Like item fit statistics, person fit statistics are summarized using the mean squares method and the *t* distribution method (Bond & Fox, 2001). Same as item fit statistics, the value of 1.00 is the projected mean square value for both infit and outfit statistics (Wright & Linacre, 1994). Wright and Linacre (1994) argued that researchers should use more flexible mean square values when they evaluate person fit statistics—note that 0.80 and 1.20 are recommended as the cutoff points by researchers for item fit (Bond & Fox, 2001; Engelhard, 2013).

Based on Wright and Linacre's (1994) argument, the lower cutoff value of 0.50 and the upper cutoff point of 1.50 were used for person fit statistics (i.e., infit mean square statistics) in the present study. These values were chosen because Linacre (2002a) argued that an infit value between 0.50 and 1.50 is "[p]roductive for measurement" (p. 878). Same as item fit statistics, mean square person fit values above 1.50 are a greater threat than mean square person fit values below 0.50. Mean square person fit values above 1.50 denote the randomness of the student responses.

Analyses of person fit in the current study examined the quality of the items in the TTCT-figural Form A as a group. An infit mean square value above 1.50 indicated that the variation in the person measure was at least 50% more than Rasch measurement theory predicted and that the

student did not respond to the items as predicted by Rasch measurement theory. An infit mean square value below 0.50 suggested that the variation in the person measure was at least 50% less than Rasch measurement theory predicted and that the student responded to the items more perfectly than Rasch measurement theory projected. An infit mean square value between 0.50 and 1.50 implied that the items had good quality and that the student responded to the items as predicted by Rasch measurement theory.

Person reliability index. Person reliability index is used to assess whether an instrument delivers item-independent (i.e., invariant) measurement of the trait levels of examinees (Bond & Fox, 2001). The person reliability index can have a value between 0 and 1.00. The cutoff point suggested for invariant measurement of the trait levels of examinees is .80 (Linacre, 2016). Any value above .80 indicates a sufficiently high possibility that invariant measurement of the trait levels of examinees is achieved.

The person reliability index indicates whether the trait levels of a group of examinees would remain at the same level if the examinees took another set of items measuring the same latent trait (Wright & Masters, 1982). The person reliability index in the present study revealed whether the creative thinking skill levels of the students would remain at the same level if another test measuring the same latent trait (e.g., the TTCT-figural Form B) were administered.

CHAPTER 4

RESULTS

The mean item scores will be summarized first. Afterwards, the findings of Rasch measurement theory analyses will be presented for all the variables measured in the TTCT-figural Form A as well as for elaboration-I. Based on the scoring method of the items for a particular variable (i.e., dichotomous or polytomous scoring), the following Rasch parameters will be summarized: item fit, item difficulty, average item difficulty, category thresholds, differential item functioning, item information, item reliability (i.e., sample-independent item calibration), and person reliability (i.e., item-independent measurement of the trait levels of the students).

Mean Item Scores

Mean item scores (and activity scores for elaboration) were calculated for all the variables. The ranges for the mean item (or activity) scores were as follows: from 0.77 to 0.99 for fluency, from 0.34 to 0.83 for originality, from 7.01to 8.46 for elaboration, from 1.04 to 1.93 for elaboration-I, from 0.28 to 1.10 for abstractness of titles, from 0.64 to 1.43 for resistance to premature closure, from 0.02 to 0.23 for emotional expressiveness, from 0.01 to 0.47 for storytelling articulateness, from 0.02 to 0.40 for movement or action, from 0.07 to 0.43 for expressiveness of titles, from 0 to 0.01 for synthesis of incomplete figures, from 0.01 to 0.03 for synthesis of lines, from 0.10 to 0.70 for unusual visualization, from 0.01 to 0.11 for internal visualization, from 0.29 to 0.50 for extending or breaking boundaries, from 0.01 to 0.12 for

humor, from 0.07 to 0.52 for richness of imagery, from 0.04 to 0.54 for colorfulness of imagery, and from 0.03 to 0.34 for fantasy. The mean item scores are reported in Table 1.

Rasch Measurement Theory Analyses

The test items (or activities for elaboration) were analyzed for all the variables measured in the TTCT-figural Form A as well as for elaboration-I. Item fit, differential item functioning, sample-independent (i.e., invariant) item calibration, and item-independent (i.e., invariant) measurement of the trait levels of the students were explored for all the variables. Item difficulty was examined for only the dichotomously scored variables. Average item difficulty and item scaling were inspected for only the polytomously scored variables. Item information was estimated for only the norm-referenced variables (i.e., fluency, originality, elaboration, abstractness of titles, and resistance to premature closure).

Fluency

Item fit, item difficulty, differential item functioning, item information, invariant item calibration, and invariant measurement of fluency skill levels of the students were examined. Infit mean square statistics showed that the overwhelming majority of the items had a good fit regarding fluency. The infit mean square values (I-MNSQ) were within the suggested range, 0.80 and 1.20 (Engelhard, 2013), for 16 of the items. Only the item in Activity 1 had an infit mean square value out of this range (I-MNSQ = 1.29). However, the standardized infit (I-STZD) value for this item was 1.10 which was within the suggested range for a good fit, -2.00 and 2.00 (Bond & Fox, 2001). The standardized infit values for the items in Activity 2 and Activity 3 were also within the recommended range. Item fit statistics are presented in Table 2.

Regarding item difficulty (δ), the items ranged from -2.72 to 1.36 logits for the whole sample. The majority of the items were clustered between -0.83 and 0.72 logit points (see the

third column in Figure 1). The first item in Activity 2 was the easiest in terms of creating a meaningful figural response, while the sixth item in Activity 3 was the most difficult. Item difficulties are reported in Table 2.

When gender was included in the model, analyses yielded different item difficulty levels for boys and girls. Differences in item difficulty ranged from 0.02 to 1.41 logits. The disparity was the largest for the first item in Activity 2 (easier for boys) and the smallest for the sixth item in Activity 3 (easier for girls). In order to test whether these differences were significant, Welch's *t*-test was employed. None of the differences in item difficulty was statistically significant at the $p < .05$ level. Findings of differential item functioning are summarized in Table 2.

Item information was estimated for each test item. Analyses yielded a bell-shaped information function for each individual item (see Figure 2). The midpoint of the information function of an item was the item's difficulty point on the latent trait continuum. Each item provided the highest amount of information (i.e., lowest standard error of measurement) on fluency at its difficulty point. The amount of information provided by a particular item decreased towards the far ends of the latent trait continuum from the item's difficulty point. As seen in Figure 2, each individual test item delivered the same amount of information on fluency. However, each item provided this information within a different logit range on the latent trait continuum due to the disparities in item difficulty. Figure 2 displays the item information functions.

There was considerable variation in the trait levels of the students. This variation led to the estimation of a considerably high item reliability index (IRIX; IRIX = .87). The item reliability index was, however, .03 below the cutoff point suggested for invariant item calibration

(Bond & Fox, 2001). A considerably low person reliability index (PRIX) was estimated for fluency (PRIX = .12). The person reliability index implied that the variation in the difficulty levels of the items was much lower than the Rasch model projected for invariant measurement of fluency skills (Bond & Fox, 2001).

Originality

Examined were item fit, item difficulty, differential item functioning, item information, invariant item calibration, and invariant measurement of originality skills. All the items were estimated to have a good fit regarding originality. The infit mean square values were between 0.80 and 1.20 for the items. Similarly, the standardized infit values were also within the recommended range for a good fit for the test items. Fit statistics are reported in Table 3.

The items were spread between -1.47 and 1.13 logit points with respect to difficulty (see the third column in Figure 4). The easiest item in terms of creating an original response was the third item in Activity 2. The most difficult item was the first item in Activity 3. Table 3 presents the findings of item difficulty.

When item difficulties were estimated for each gender, unequal difficulty levels were obtained for boys and girls. Differences in item difficulty were between 0.02 and 0.58 logits. The largest disparity was estimated for the first item in Activity 2 (easier for boys), and the smallest difference was obtained for the sixth item in the same activity (easier for girls). Welch's *t*-test analyses revealed that none of the differences in item difficulty was statistically significant at the $p < .05$ level. Table 3 summarizes the differential item functioning analyses.

Bell-shaped information functions were obtained for the test items (see Figure 5). The amount of information provided by a particular item on originality was the highest at the item's difficulty point on the latent trait continuum—the standard error of measurement was the lowest

at this point. The amount of item information declined in the direction of the far ends of the latent trait continuum from the item's difficulty point. Each individual item delivered the same amount of information on originality. However, due to the variation in item difficulty, each item provided this information at a different portion of the latent trait continuum. Information functions of the items are presented in Figure 5.

The variation in the trait levels of the students was sufficiently high. An item reliability index value of .95, which was above the cutoff point for invariant item calibration, was estimated. The items were not spread well on the latent trait continuum with respect to difficulty. Hence, a moderately high person reliability index was estimated (PRIX = .34). The person reliability index implied that invariant measurement of originality skills was not achieved.

Elaboration

The activities were analyzed for fit, average difficulty, scaling, differential item (i.e., activity) functioning, information, invariant calibration of the scaling, and invariant measurement of elaboration skills. The infit mean square and standardized infit values were within the suggested ranges for a good fit for all three activities. The infit mean square value for Activity 1 was 0.81, Activity 2 was 0.93, and Activity 3 was 1.20. The standardized infit value of -1.40 was estimated for Activity 1, -0.40 was estimated for Activity 2, and 1.40 was estimated for Activity 3.

The activities ranged from -0.83 to 0.86 logits regarding average difficulty (see the third column in Figure 7). Activity 1 was the easiest in terms of elaboration, while Activity 2 was the most difficult. The average difficulty level of Activity 3 was -0.04 logits. The Rasch-Andrich thresholds estimated for Activity 1 were -4.00, -1.97, -0.53, -0.69, and 1.05 logit points from the first threshold to the last. For Activity 2, -2.31, -0.28, 2.16, 2.00, and 2.74 were the threshold

values. The threshold values obtained for Activity 3 were -3.21, -1.18, 1.26, 1.10, and 1.84 logits. The average difficulty levels of the activities are presented in Table 4.

The average difficulty levels of the activities were different for each gender. Disparities in average difficulty ranged from 0.19 to 0.31 logits. The disparity was the largest for Activity 1 (easier for boys) and the smallest for Activity 3 (easier for boys). The difference for Activity 2 was 0.28 logits; the activity was easier for boys. None of the differences was statistically significant at the $p < .05$ level. Analyses of differential item (i.e., activity) functioning are reported in Table 4.

Each information function was estimated to have a wide peak (see Figure 9). The height of a peak increased towards the right end of the information function and then declined. The information function of a particular activity also declined as the function approached the left end of the latent trait continuum. Each activity delivered the highest amount of information (i.e., lowest standard error of measurement) on elaboration on the right side of the information function. Although all three activities provided the same amount of information on elaboration, each activity delivered this information within a different logit range due to the differences in average difficulty. Information functions are illustrated in Figure 9.

The variation in the elaboration skills of the students was high enough for invariant measurement of the scaling of the activities. The item (i.e., activity) reliability index was estimated to be .97. Because there were three activities and the variation in the difficulty levels of the activities as well as the response options was not high enough, a value of .39 was estimated for the person reliability index.

Elaboration-I

Item fit, average item difficulty, item scaling, differential item functioning, invariant item calibration, and invariant measurement of elaboration-I skills were investigated for the items in Activity 2 and for the first six items in Activity 3. The overwhelming majority of the items had a good fit. The infit mean square values were within the recommended range for 15 of the items. Only the second item in Activity 3 showed overfit (I-MNSQ = 0.75). The standardized infit value for this item also denoted overfit (I-STZD = -2.40). The rest of the standardized infit values were between -2.00 and 2.00. Table 5 shows the fit statistics.

The items were clustered between -0.22 and 0.25 logit points regarding average item difficulty (see the third column in Figure 11). The easiest item regarding item-level elaboration was the second item in Activity 3. The most difficult item, on the other hand, was the tenth item in Activity 2. For a hypothetical item with the average item difficulty of 0, the Rasch-Andrich thresholds were -0.14, -0.20, -0.03, 0.36, 0.41, and -0.40 logit points (The Rasch-Andrich thresholds can be estimated for the actual items on the test by summing the average difficulty level of an item and the value for a particular threshold. For instance, for the first item in Activity 2 [δ = -0.03], the Rasch-Andrich thresholds were -0.17, -0.23, -0.06, 0.33, 0.38, and -0.43 logit points). The average item difficulties are presented in Table 5.

Gender-based item analyses estimated different average item difficulty values for boys and girls. Average item difficulty differences were between 0.01 and 0.19 logits. The largest disparity was obtained for the sixth item in Activity 3 (easier for girls), and the smallest difference was estimated for the eighth item in Activity 2 (easier for boys). None of the differences was, however, statistically significant at the $p < .05$ level according to Welch's *t*-test results. Analyses of differential item functioning are summarized in Table 5.

The variation in the trait levels of the students was not high enough to provide invariant calibration of the items. A considerably high item reliability index was obtained for elaboration-I (IRIX = .76), but it was below the suggested cutoff value. There was considerable variation in the average difficulty levels of the items as well as the response options, but the variation was not sufficiently high to provide invariant measurement of elaboration-I skill levels of the students. The person reliability index was estimated to be .75.

Abstractness of Titles

Analyses were conducted to examine item fit, average item difficulty, item scaling, differential item functioning, item information, invariant item calibration, and invariant measurement of abstractness of titles skills. The item in Activity 1 and all 10 items in Activity 2 were analyzed. Fit statistics indicated that the majority of the items had a good fit regarding abstractness of titles. Only the ninth item (I-MNSQ = 0.79) and the tenth item (I-MNSQ = 1.31) in Activity 2 showed misfit. The standardized infit values were between -2.00 and 2.00 for all the test items, including the two misfitting items. Fit statistics are summarized in Table 6.

The lower logit point of -1.89 and the upper logit point of 0.41 were estimated for average item difficulty. The majority of the items were located between -0.21 and 0.41 logit points (see the third column in Figure 14). The item in Activity 1 was estimated to be the easiest in terms of abstractness of titles, while the eighth item in Activity 2 was the most difficult. For a hypothetical item with the average item difficulty of 0, the Rasch-Andrich thresholds were -1.74, -1.90, and 3.63 logit points (The Rasch-Andrich thresholds can be estimated for the actual items on the test by summing the average difficulty level of an item and the value for a particular threshold. For instance, for the first item in Activity 2 [$\delta = 0.08$], the Rasch-Andrich thresholds were -1.66, -1.82, and 3.71 logit points). Table 6 presents the average item difficulties.

The rating scale model estimated unequal average item difficulty levels for each gender. Differences in average item difficulty ranged from 0.05 to 0.98 logits. The disparity was the largest for the second item in Activity 2 (easier for girls) and the smallest for the fourth item in the same activity (easier for girls). According to Welch's *t*-test results, the disparity in average item difficulty was statistically significant for the second item ($t(113) = -2.73$, $p = .01$; easier for girls) and the ninth item ($t(129) = 2.32$, $p = .02$; easier for boys) in Activity 2. Analyses of differential item functioning are reported in Table 6.

The information function of a particular item had two peaks (see Figure 16). One peak was higher on the right side of the information function, while the other peak was higher on the left side of the information function. However, the peak on the left side was higher, meaning that the amount of standard error of measurement was lower for the first three response options. Each individual test item provided the same amount of information on abstractness of titles; however, because the items differed from each other with respect to average difficulty, each item delivered this information within a different logit range on the latent trait continuum. Figure 16 presents the information functions of the items.

The sample of the study was spread well regarding abstractness of titles skill levels. Therefore, a sufficiently high value for the item reliability index was estimated (IRIX = .94). On the other hand, the person reliability index was moderately high (PRIX = .56), implying that the variation in the difficulty levels of the items as well as the response options was not as much as the rating scale model predicted.

Resistance to Premature Closure

Examined were item fit, average item difficulty, item scaling, differential item functioning, item information, invariant item calibration, and invariant measurement of

resistance to premature closure skills. Only the items in Activity 2 were analyzed. Item fit analyses revealed that the items had a good fit regarding resistance to premature closure. Both the infit mean square and standardized infit values were within the suggested ranges for a good fit for all the items in Activity 2. Fit statistics are reported in Table 7.

The items were located between -0.73 and 0.71 logit points on the latent trait continuum in terms of average item difficulty; all the items were between -0.73 and 0.71 logits (see the third column in Figure 18). The easiest item was the fourth item, and the most difficult item was the fifth item. For a hypothetical item with the average item difficulty of 0, the Rasch-Andrich thresholds were 0.46 and -0.46 logit points (The Rasch-Andrich thresholds can be estimated for the actual items on the test by summing the average difficulty level of an item and the value for a particular threshold. For instance, for the first item in the activity [$\delta = 0.15$], the Rasch-Andrich thresholds values were 0.61 and -0.31 logits). Average item difficulties are presented in Table 7.

Analyses yielded different average item difficulties for each gender. Average item difficulty differences ranged from 0.03 to 0.50 logits. The largest disparity was obtained for the seventh item (easier for boys), and the smallest difference was estimated for the third item (easier for boys) as well as the tenth item (easier for boys). Welch's *t*-test results showed that the seventh item was significantly easier for boys ($t(152) = 2.41$, $p = .02$). Table 7 summarizes the findings of differential item functioning.

A bell-shaped information function was estimated for each test item (see Figure 20). The amount of information provided by a particular item on resistance to premature closure was the highest at the item's average difficulty point on the latent trait continuum—the standard error of measurement was the lowest at this point. The amount of item information decreased in the direction of the far ends of the latent trait continuum from the item's average difficulty point.

Each individual test item delivered the same amount of information on resistance to premature closure. Due to the differences in average item difficulty, however, each item delivered this information within a different logit range on the latent trait continuum. Item information functions are illustrated in Figure 20.

Analyses of invariant item calibration showed that the item reliability index was .94. This value indicated that the variation in the resistance to premature closure skill levels of the students was sufficiently high. The person reliability index was moderately high (PRIX = .65); it implied that the variation in item difficulty did not yield invariant measurement of resistance to premature closure skills.

Emotional Expressiveness

The items were analyzed for item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of emotional expressiveness skills. The infit mean square values were between 0.80 and 1.20 for all the test items except three items: the fifth (I-MNSQ = 1.21), sixth (I-MNSQ = 0.77) and seventh (I-MNSQ = 0.73) items in Activity 2. However, the standardized infit values for the test items, including these three misfitting items, were within the recommended range. Table 8 summarizes the fit statistics.

The items were spread between -2.59 and 1.22 logit points with respect to difficulty. The majority of the items were between -0.40 and 1.22 logits (see the third column in Figure 22). The item in Activity 1 was the easiest in terms of expressing an emotion in a response. The fourth item in Activity 3, on the other hand, was the most difficult. Item difficulties are reported in Table 8.

The Rasch model estimated different item difficulty levels for each gender. Differences in item difficulty ranged from 0.16 to 1.76 logits. The disparity was the largest for the fourth item

in Activity 3 (easier for boys) and the smallest for the first item as well as the second item in Activity 3 (both easier for boys). Welch's *t*-test results revealed that the difficulty levels were significantly disparate for the fourth item ($t(66) = 2.02$, $p = .04$; easier for boys) as well as the fifth item ($t(53) = -2.22$, $p = .03$; easier for girls) in Activity 2. Analyses of differential item functioning are presented in Table 8.

There was considerable variation in the trait levels of the students, but it was not as much as the Rasch model predicted. Thus, an item reliability index of .81, which was .09 smaller than the minimum recommended value for invariant item calibration, was estimated. Because the items lacked sufficient variation in difficulty, the person reliability index was estimated to be zero.

Storytelling Articulateness

Item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of storytelling articulateness skills were investigated. The infit mean square values were within the recommended range for a good fit for all the items except the item in Activity 1 (I-MNSQ = 1.22) and the third item in Activity 3 (I-MNSQ = 0.71). The standardized infit values were, however, between -2.00 and 2.00 for all the items. Fit statistics are presented in Table 9.

The difficulty levels of the items ranged from -3.95 to 2.16 logits, with the majority of the items being between -0.79 and 0.75 logits (see the third column in Figure 24). The easiest item in terms of creating a figural response depicting a story was the item in Activity 1, while the most difficult item was the tenth item in Activity 2. Analyses of item difficulty are summarized in Table 9.

Item analyses estimated unequal item difficulties for each gender. Item difficulty differences ranged from 0.02 to 1.74 logits. The largest disparity was estimated for the seventh item in Activity 2 (easier for boys), and the smallest difference was obtained for the second item in the same activity (easier for boys). According to Welch's *t*-test results, the seventh item in Activity 2 was significantly easier for boys ($t(98) = 2.27$, $p = .03$). Differential item functioning analyses are reported in Table 9.

The variation in the trait levels of the students was considerably high; yet the item reliability index (IRIX = .89) was .01 lower than the cutoff value suggested for invariant calibration of the items. The person reliability index, on the other hand, was zero due to the lack of sufficient variation in the difficulty levels of the items.

Movement or Action

Analyses were conducted to examine item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of movement or action skills. Infit mean square statistics showed that the test items had a good fit regarding movement or action. The infit mean square values were between 0.80 and 1.20 for all the items. The standardized infit values also denoted a good fit. All the standardized infit values were within the recommended range. Fit statistics are reported in Table 10.

The lower logit point of -2.53 and the upper logit point of 1.82 were estimated for item difficulty. The majority of the items were located between -1.26 and 0.49 logit points (see the third column in Figure 26). The item in Activity 1 was estimated to be the easiest in terms of expressing movement or action in a response. The most difficult item was the fifth item in Activity 2. Item difficulties are presented in Table 10.

Gender-based item analyses estimated different item difficulty levels for each gender. Differences in item difficulty were between 0.05 and 1.62 logits. The disparity was the largest for the sixth item in Activity 3 (easier for boys) and the smallest for the first item in Activity 2 (easier for girls). Welch's *t*-test results showed that the disparity in item difficulty was statistically significant for the ninth item in Activity 2 ($t(110) = -2.11$, $p = .04$) which was easier for girls. Table 10 summarizes the findings of differential item functioning.

Analyses of invariant item calibration showed that there was considerable variation in the trait levels of the students. However, the variation led to the estimation of an item reliability index value of .89, which was .01 below the cutoff point recommended for invariant item calibration. Because the items were clustered within a narrow range with respect to difficulty regarding movement or action, the person reliability index value of 0 was obtained.

Expressiveness of Titles

Item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of expressiveness of titles skills were explored. The infit mean square values were within the recommended range for a good fit for all the items with the exception of the item in Activity 1 (I-MNSQ = 1.35) and the first item (I-MNSQ = 0.74) as well as the second item (I-MNSQ = 1.31) in Activity 2. Similarly, the overwhelming majority of the standardized infit values were between -2.00 and 2.00. Only the standardized infit value for the item in Activity 1 showed underfit (I-STZD = 3.60). Table 11 presents the fit statistics.

The items were located between -3.15 and 0.90 logit points with respect to item difficulty. The majority of the items were clustered between -0.35 and 0.90 logit points (see the third column in Figure 28). The item in Activity 1 was estimated to be the easiest in terms of

giving an expressive title to a figural response. The fifth item in Activity 2, on the other hand, was the most difficult. Item difficulties are reported in Table 11.

Difficulty levels of the items were different for each gender. Differences in item difficulty ranged from 0.05 to 1.94 logits. The largest disparity was obtained for the third item in Activity 3 (easier for boys), and the smallest difference was estimated for the sixth item in Activity 2 (easier for boys). Welch's *t*-test results revealed that the disparities in item difficulty were statistically significant for the following three items: the seventh item in Activity 2 ($t(94) = 2.04$, $p = .04$; easier for boys), and the first item ($t(66) = -2.35$, $p = .02$; easier for girls) as well as the third item ($t(85) = 2.72$, $p = .01$; easier for boys) in Activity 3. Differential item functioning analyses are summarized in Table 11.

The item reliability index was estimated to be .87, which was .03 below the suggested value for invariant item calibration. In other words, the variation in the trait levels of the students was a little lower than the Rasch model projected. There was some degree of variation in the difficulty levels of the items, but the variation was not high enough for invariant measurement of expressiveness of titles skills. Thus, a relatively low person reliability index was estimated (PRIX = .24).

Synthesis of Incomplete Figures

Rasch measurement theory analyses were conducted to examine item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of synthesis of incomplete figures skills. However, because only two of the students participating in the study combined two or more incomplete figures in Activity 2 into one response, no meaningful results were obtained for the items regarding synthesis of incomplete figures. Results indicated that synthesis of incomplete figures was rarely expressed by second graders. A bigger sample size is

needed to obtain consistent response patterns and to analyze the items regarding synthesis of incomplete figures.

Synthesis of Lines

The first six items in Activity 3 were analyzed for item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of synthesis of lines skills. However, only a few (six) of the students participating in the study combined two or more parallel line pairs into one response. Therefore, the analyses did not provide meaningful results for the items regarding synthesis of lines. Like synthesis of incomplete figures, synthesis of lines was not a skill that was frequently manifested by second graders. A larger sample size is needed for conducting item-level analyses and obtaining interpretable results.

Unusual Visualization

Analyses were conducted to examine item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of unusual visualization skills. The items were estimated to have a good fit regarding unusual visualization. The infit mean square values were between 0.80 and 1.20 for all the items. Good item fit was also implied by the standardized infit values. All the standardized infit values were within the suggested range. Table 12 summarizes the fit statistics.

The items were spread between -2.27 and 1.11 logit points with respect to difficulty (see the third column in Figure 30). The easiest item in terms of creating a figural response with unusual perspective was the item in Activity 1, while the most difficult item was the tenth item in Activity 2. Findings of item difficulty are reported in Table 12.

The Rasch model estimated unequal item difficulty levels for each gender. Differences in item difficulty ranged from 0 to 0.79 logits. The difference was the largest for the tenth item in

Activity 2 (easier for girls) and zero for the second item in the same activity. None of the differences in item difficulty was statistically significant at the $p < .05$ level. Analyses of differential item functioning are reported in Table 12.

Because the variation in the trait levels of the students was sufficiently high, an item reliability index over .90 was estimated (IRIX = .94). This value implied that the requirement for invariant item calibration was fulfilled. There was some degree of variation in the difficulty levels of the items, but it was not as much as the Rasch model projected. Therefore, a value of .15 was estimated for the person reliability index, indicating that invariant measurement of unusual visualization skills was not achieved.

Internal Visualization

Item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of internal visualization skills were examined. Both the infit mean square and standardized infit statistics indicated that all the items had a good fit regarding internal visualization. The infit mean square values were between 0.80 and 1.20, and the standardized infit values were between -2.00 and 2.00. Table 13 summarizes the fit statistics.

The items ranged from -0.93 to 2.22 logits with respect to item difficulty, with the majority of the items being between -0.93 and 0.51 logits (see the third column in Figure 32). The item in Activity 1 was the easiest in terms of creating a figural response with internal visualization. The most difficult item, on the other hand, was the tenth item in Activity 2. Item difficulties are presented in Table 13.

Item analyses estimated unequal item difficulties for each gender. Differences in item difficulty were between 0.10 and 2.39 logits. The largest disparity was obtained for the second item in Activity 2 (easier for boys), and the smallest difference was estimated for the sixth item

the same activity (easier for boys). Welch's *t*-test results showed that the disparity in item difficulty was statistically significant for the second item in Activity 2 ($t(77) = 2.18$, $p = .03$), which was easier for boys. Differential item functioning analyses are reported in Table 13.

Although there was some degree of variation in the trait levels of the students, the variation was not high enough to provide invariant calibration of the items. Therefore, the value estimated for the item reliability index (IRIX = .65) was lower than the minimum suggested cutoff point. Due to the lack of sufficient variation in item difficulty, the person reliability index was estimated to be zero.

Extending or Breaking Boundaries

Examined were item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of extending or breaking boundaries skills. The first six items in Activity 3 were analyzed for extending or breaking boundaries. The items had a good fit regarding extending or breaking boundaries. The infit mean square values were within the recommended range for a good fit for all six items. The standardized infit values were also within the suggested range for a good fit. Fit statistics are reported in Table 14.

The lower logit point of -0.71 and the upper logit point of 0.69 were estimated for item difficulty; all the items were located between these logit points (see the third column in Figure 34). The third item was estimated to be the easiest in terms of extending or breaking the boundaries of parallel lines. The sixth item was the most difficult. Item difficulties are summarized in Table 14.

Analyses yielded different item difficulty levels for each gender. Item difficulty differences ranged from 0.13 to 1.08 logits, with the disparity being the largest for the sixth item (easier for girls) and the smallest for the first item (easier for boys). Welch's *t*-test results

revealed that the sixth item was significantly easier for girls ($t(85) = -2.13$, $p = .03$). Analyses of differential item functioning are presented in Table 14.

Analyses of invariant item calibration showed that the trait levels of the students had sufficient variation to provide invariant calibration of the items. The item reliability index was estimated to be .90. There was some variation in item difficulty, but the variation was not as much as the Rasch model predicted. Thus, a moderately high person reliability index was estimated (PRIX = .41), indicating that invariant measurement of extending or breaking boundaries skills was not achieved.

Humor

The items were analyzed for item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of humor skills. The infit mean square values were within the suggested range for a good fit for all the items with the exception of two items with underfit: the item in Activity 1 (I-MNSQ = 1.45) and the ninth item in Activity 2 (I-MNSQ = 1.21). The standardized infit value for the item in Activity 1 (I-STZD = 3.10) also denoted underfit. The rest of the items had standardized infit values between -2.00 and 2.00. Fit statistics are presented in Table 15.

The items were spread between -1.73 and 1.11 logit points with respect to item difficulty (see the third column in Figure 36). The easiest item in terms of creating a humorous response was the item in Activity 1, while the most difficult item was the second item in Activity 3. Item difficulty analyses are summarized in Table 15.

The Rasch model yielded unequal item difficulty levels for boys and girls. Differences in item difficulty ranged from 0.13 to 1.78 logits. The largest disparity was estimated for the fifth item in Activity 2 (easier for girls), and the smallest difference was obtained for the first item in

Activity 3 (easier for girls). None of the differences in item difficulty was, however, statistically significant at the $p < .05$ level. Findings of differential item functioning are reported in Table 15.

Analyses of invariant item calibration yielded an item reliability index value of .46. This value suggested that the variation in the trait levels of the students was not high enough to provide invariant item calibration. Because of the lack of sufficient variation in the difficulty levels of the items, the value for the person reliability index was estimated to be zero.

Richness of Imagery

Item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of richness of imagery skills were explored. Fit statistics showed that the items had a good fit regarding richness of imagery. Both the infit mean square and standardized infit values were within the suggested ranges for all the items. Fit statistics are reported in Table 16.

Item difficulties ranged from -2.38 to 0.97 logits. The majority of the items were clustered between -0.73 and 0.97 logit points (see the third column in Figure 38). The item in Activity 1 was estimated to be the easiest in terms of creating a memorable figural response. The most difficult item was the sixth item in Activity 3. Item difficulties are presented in Table 16.

Gender-based analyses yielded different item difficulty levels for boys and girls. Differences in item difficulty were between 0.08 and 1.24 logits. The disparity was the largest for the first item in Activity 3 (easier for boys) and the smallest for the item in Activity 1 (easier for girls). Welch's *t*-test analyses showed that none of the differences in item difficulty was statistically significant at the $p < .05$ level. Table 16 summarizes the analyses of differential item functioning.

The variation in the trait levels of the students was considerably high. However, this variation did not yield an item reliability index over .90. The item reliability index was .89. There was some variation in the difficulty levels of the items, but it was not high enough for invariant measurement of richness of imagery skills. The value of .20 was obtained for the person reliability index.

Colorfulness of Imagery

Analyses were conducted to examine item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of colorfulness of imagery skills. The infit mean square values were between 0.80 and 1.20 for all the items, meaning that the items had a good fit regarding colorfulness of imagery. The standardized infit values supported these finding. The standardized infit values were within the recommended range for all the test items. Table 17 summarizes the fit statistics.

The difficulty levels of the items ranged from -2.72 to 1.20 logits, with the majority of the items being between -0.30 and 0.49 logits (see the third column in Figure 40). The item in Activity 1 was the easiest in terms of creating an exciting figural response, while the first item in Activity 3 was the most difficult. Item difficulties are presented in Table 17.

Analyses yielded different item difficulties for each gender. The range for the differences in item difficulty was from 0 to 0.95 logits. The largest disparity was obtained for the ninth item in Activity 2 (easier for boys), and zero difference was estimated for the fourth item in Activity 3. None of the differences in item difficulty was statistically significant at the $p < .05$ level. Differential item functioning analyses are reported in Table 17.

Due to the sufficient variation in the trait levels of the students, the value for the item reliability index was estimated to be .91. On the other hand, because the variation in item

difficulty was not as much as the Rasch model projected, the person reliability index was below the recommended value (PRIX = .11). This indicated that invariant measurement of colorfulness of imagery skills was not obtained.

Fantasy

Examined were item fit, item difficulty, differential item functioning, invariant item calibration, and invariant measurement of fantasy skills. The infit mean square values were within the suggested range for a good fit for 13 of the items on the test. The following four items showed misfit: the sixth item (I-MNSQ = 0.79) as well as the ninth item (I-MNSQ = 1.22) in Activity 2, and the second item (I-MNSQ 0.78) as well as the sixth item (I-MNSQ 0.79) in Activity 3. However, the standardized infit values for all the test items, including these four misfitting items, were between -2.00 and 2.00. Fit statistics are presented in Table 18.

The lower logit point of -3.13 and the upper logit point of 0.99 were estimated for item difficulty. The majority of the items were located between -0.93 and 0.99 logit points (see the third column in Figure 42). The easiest item in terms of expressing fantasy in a response was the item in Activity 1. The most difficult item, on the other hand, was the first item in Activity 3. Table 18 reports the findings of item difficulty.

Gender-based item analyses estimated unequal item difficulty levels for boys and girls. Differences in item difficulty ranged from 0.04 to 1.93 logits, with the disparity being the largest for the fifth item in Activity 2 (easier for girls) and the smallest for the seventh item in the same activity (easier for boys). According to Welch's *t*-test results, none of the differences in item difficulty was statistically significant at the $p < .05$ level. Table 18 summarizes the analyses of differential item functioning.

Although there was considerable variation in the trait levels of the students, the variation was not as much as the Rasch model projected. Therefore, an item reliability index below the recommended value for invariant item calibration was estimated (IRIX = .83). Because of the lack of sufficient variation in item difficulty, the person reliability index was estimated to be zero.

CHAPTER 5

DISCUSSION

Discussion will be presented separately for each individual variable measured in the TTCT-figural Form A as well as for elaboration-I. For each variable, infit mean square and standardized infit statistics will be discussed first because Linacre (2004) opined that fit statistics were good starting points to inspect Rasch analyses. After infit statistics are discussed, findings of item difficulty, average item difficulty, item scaling, differential item functioning, item information, invariant item calibration, and invariant measurement of the trait levels of the students will be critiqued.

Fluency

Four assumptions are currently made about the test items regarding fluency. The items are assumed to have a good fit, to be equally difficult, to behave the same for students of each gender, and to involve the same amount of standard error of measurement, which is at the same level along the latent trait continuum. In order to test these assumptions, the dichotomous Rasch model was utilized.

Item Fit

The infit mean square values were between 0.80 and 1.20 for 16 of the items. Only the item in Activity 1 had an infit mean square value out of this range (I-MNSQ = 1.29). The infit mean square values for 16 good fitting items showed that the variation in the item measures was not considerably larger or smaller than the Rasch model predicted. This finding suggested that the responses given to each of these 16 items were affected by one factor and independent of one

another. The standardized infit values for these 16 items supported this conclusion. The standardized infit values for these items were between -2.00 and 2.00, and they implied that a good fit for each of these 16 items was highly possible.

The infit mean square value for the item in Activity 1 (I-MNSQ = 1.29) indicated that there was 29% more variation in the item measures than the Rasch model predicted. This variation denoted the haphazardness in the responses given to this item and implied that a second factor, other than fluency, might have impacted the student responses. The standardized infit value for the item in Activity 1 (I-STZD = 1.10), however, suggested that the existence of a second factor was not likely. Additionally, this 29% more variation can be neglected because Adams and Khoo (1996; as cited in Wilson, 2005, p. 129) argued that up to 33% more variation in the item measures than Rasch measurement theory predicted was not large enough to denote a second factor.

The study showed that there was acceptable variation in the item measures for 16 of the items. The item in Activity 1 was estimated to involve 29% more variation than the Rasch model predicted; however, not only was this variation negligible but also misfit was not likely for this item (as implied by the standardized infit value). In other words, the variation in the responses given to the item in Activity 1 was not detrimental. Findings of the study confirmed the assumption about item fit for 16 items and implied that the item in Activity 1 had an acceptable fit.

Item Difficulty

The items ranged from -2.72 to 1.36 logits with respect to difficulty. The majority of the items were clustered between -0.83 and 0.72 logit points. Activity 1 contained the third easiest item (δ = -0.81) regarding creating a meaningful figural response. The items in Activity 2, on the

other hand, were either below or above the 0 logit point—the easiest item on the test was in Activity 2 (the first item). Similarly, the first six items in Activity 3 were either below or above the 0 logit point—the most difficult item on the test was in this activity (the sixth item).

It was unexpected to observe the variation in the difficulty levels of the items in Activity 3 because they are essentially the same stimulus—there was a 1.70-logit difference between the easiest item and the most difficult item in the activity. A small degree of variation would be negligible, but a 1.70-logit difference is eye-catching. The large disparity between the easiest item and the most difficult item in the activity might be due to the missing observations for these two items. The noticeable relation between the missing observations and item difficulty supported this inference. There were 53 missing observations for the sixth item (highest in the activity), and it was estimated to be the most difficult item in the activity. On the other hand, there were 35 missing observations for the first item and 36 missing observations for the second item (two lowest in the activity), and they were the two easiest items in the activity.

Results indicated that some items were more challenging (and *vice-versa*) than others in terms of creating mental pictures from them and converting a particular item (i.e., stimulus) into a meaningful figural response. The difficulty level of an item may be impacted by the student's familiarity with the shape of the stimulus. It is likely that the stimuli that are perceived more often than others on objects are relatively easier with respect to creating figural responses from them within a short amount of time—when there is unlimited time, all the stimuli are expected to be equally challenging (i.e., difficult) from a statistical point of view.

In order to determine whether the difficulty levels of the items differed from each other, strata statistics were utilized. Strata statistics are recommended to be used when the extreme responses in the data represent actual performance levels (Linacre, 2013; Wright & Masters,

2002), as did they in the present study—there were several students with a perfect fluency score. The strata value of 3.86 was estimated. This value indicated that there were three distinct item groups in the TTCT-figural Form A with respect to item difficulty. The item reliability index (IRIX = .87) denoted a considerably—but not sufficiently—high probability that the locations of the items on the variable map (Figure 1) were determined accurately and that the three item groups existed. Findings of the study did not support the assumption that the items are equally difficult regarding fluency.

Differential Item Functioning

The Rasch model provided different item difficulty levels for boys and girls. Differences in item difficulty ranged from 0.02 to 1.41 logits. However, Welch's *t*-test results showed that none of the differences in item difficulty was statistically significant at the $p < .05$ level. This finding implied that the items behaved the same for both genders. The present study confirmed the assumption that the items in the TTCT-figural Form A do not favor students of one gender over students of the opposite gender.

Item Information

Estimations yielded bell-shaped information functions for the test items (see Figure 2). The midpoint of the information function of a particular item was the item's difficulty point on the latent trait continuum. Each item provided the highest amount of information on fluency (i.e., involved the lowest amount of standard error of measurement) at this difficulty point. The bell-shaped information functions implied that the amount of information an item delivered on fluency decreased towards the far ends of the latent trait continuum from the item's difficulty point. All the information functions offered the same amount of information. However, because the difficulty levels of the items varied, each item provided this information within a different

logit range on the latent trait continuum (see Figure 2). These findings were, however, expected due to the dichotomous scoring of the items for fluency. Under Rasch measurement theory, dichotomously scored items lead to the estimation of bell-shaped information functions, each of which has the peak at an item's difficulty point on the latent trait continuum.

Findings of the study showed that all the items provided the same amount of information on fluency and that the amount of information an item delivered on fluency (i.e., the amount of standard error of measurement involved in an item) was not at the same level along the latent trait continuum. Consequently, the assumption of the standard error of measurement being the same for all scores on fluency (i.e., all fluency scores being equally reliable) was not confirmed. The test information function estimated with 17 items supported this inference. As seen in Figure 44, the test information function for fluency also had a bell-shape with the peak at the 0 logit point, meaning that average scores on fluency involved the lowest amount of standard error of measurement (i.e., average fluency scores were the most reliable). This finding implied that when the test is administered under the standard testing conditions, some fluency scores will be more reliable.

Interim Conclusion

Four assumptions about the test items regarding fluency were examined in the present study. Findings of the study provided supporting evidence for the assumptions about item fit and differential item functioning. The assumptions about item difficulty and the standard error of measurement, however, did not receive statistical support. The item reliability index (IRIX = .87) indicated a high possibility that invariant item calibration was achieved. The majority of the item locations on the variable map (Figure 1) will not change if another group of second graders are administered the test. Because the value for the item reliability index was .03 below the

recommended value for invariant item calibration (.90; Linacre, 2016), the relative difficulty levels of a few items may change.

Results indicated that each individual test item possessed sufficient quality for measuring fluency and providing appropriate person measures for both genders. The items also had good quality as a group for the majority of the students—person fit values were not estimated for approximately 60 students because they had a perfect fluency score. All the person fit values estimated for the rest of the students were between 0.50 and 1.50 (see Figure 3). This indicated that these students responded to the items as predicted by the Rasch model.

A considerably low value for the person reliability index was estimated (PRIX = .12). This was due to the low amount of variation in the difficulty levels of the items. The value for the person reliability index implied that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 1) and that the test (with 17 items and without the time limit) did not provide invariant measurement of the fluency skill levels of the students. This finding suggested that the majority of the student locations on the variable map would change if the students were given the TTCT-figural Form B under the same testing conditions. The low amount of variation—in fact no variation—in item difficulty is, however, preferable for the TTCT-figural Form A because it is a timed test.

Originality

The following four assumptions are currently made about the items regarding originality: The items have a good fit, the items are equally difficult, the items behave the same for boys and girls, and the items involve the same amount of standard error of measurement, which is at the same level along the latent trait continuum. The dichotomous Rasch model was used to examine these assumptions.

Item Fit

The infit mean square values were between 0.80 and 1.20 for all the items. Because infit mean square values within this range indicate that the variation in the item measures was not harmfully more or less than the Rasch model projected (Bond & Fox, 2001), the study showed that the variation in the item measures was acceptable for each individual item. Put differently, a single factor affected the responses given to the items, and the responses given to an item were independent of the responses given to another item—after the latent trait (i.e., originality) was controlled. The standardized infit statistics supported these conclusions. The standardized infit values were between -2.00 and 2.00, denoting a high probability of a good fit for the items. The present study provided supporting evidence for the assumption that the items have a good fit with regard to originality.

Item Difficulty

The items were spread between -1.47 and 1.13 logit points with respect to difficulty. Activity 1 contained the second easiest item ($\delta = -1.41$) regarding creating an original figural response. The items in Activity 2, on the other hand, were either below or above the 0 logit point. In other words, some stimuli in Activity 2 (e.g., the third item; $\delta = -1.47$) generally induced original ideas, while some other stimuli (e.g., the tenth item; $\delta = 0.54$) usually led the students to generate conventional ideas. Finally, the first six items in Activity 3 were all above the 0 logit point.

This finding on the items in Activity 3 was-expected because the first responses an examinee can think of are usually conventional ideas, and the examinee can produce more original ideas as he or she continues (Acar, 2013; Runco, 1986; Ward, 1969). Because the students generated a low number of original ideas for the first six items, these items were

estimated to be relatively difficult. It is likely that the items in Activity 3 become easier in terms of creating an original response as the item number increases because examinees start to move away from the ordinary responses as time goes on.

In order to determine whether the items were at the same difficulty level, strata statistics were used. The strata value of 6.07 was estimated. This value suggested that there were six distinct item groups on the test with respect to difficulty. The sufficiently high item reliability index (IRIX = .95) supported the existence of these item groups and the accurate determination of the item locations on the variable map (Figure 4). The assumption of the items being equally difficult regarding originality was not supported.

Differential Item Functioning

The Rasch model provided unequal item difficulty levels for boys and girls. Differences in item difficulty ranged from 0.02 to 0.58 logits. Welch's *t*-test results revealed that none of the differences in item difficulty was statistically significant at the $p < .05$ level. Findings showed that the items in the TTCT-figural Form A did not favor students of one gender over students of the opposite gender in terms of creating original figural responses. The study confirmed the assumption that the items behave the same for both genders.

Item Information

Each test item was estimated to have a bell-shaped information function due to the dichotomous scoring (see Figure 5). The midpoint of this information function was the item's difficulty point, and it was where the item delivered the highest amount of information on originality (i.e., involved the lowest amount of standard error of measurement). The amount of information an item provided on originality declined towards the far ends of the latent trait continuum from the item's difficulty point. All the items had the same information function

regarding its shape, but the range of each information function corresponded to a different portion of the latent trait continuum due to the variation in item difficulty.

The present study revealed that each individual item offered the same amount of information and that the amount of information an item provided on originality (i.e., the amount of standard error of measurement involved in an item) was not at the same level along the latent trait continuum. These findings implied that the assumption of the standard error of measurement being the same for all originality scores was not true. The estimated bell-shaped test information function, which showed that the average scores on originality involved the lowest amount of standard error of measurement, supported this conclusion (see Figure 44). The test information was estimated with 17 items in the present study, but it suggested that under the standard testing conditions, some originality scores are more reliable.

Interim Conclusion

The present study investigated four assumptions about the items regarding originality. Findings provided supporting evidence for the assumptions about item fit and differential item functioning. The assumptions about item difficulty and item information were not confirmed. The item reliability index (IRIX = .95) indicated that the items were calibrated in a sample-independent manner. Therefore, the difficulty levels of the items (i.e., the item locations on the variable map [Figure 4]) are expected to remain at the same level if another group of second graders are administered the test.

Analyses demonstrated that each individual item possessed sufficient quality for measuring originality and providing appropriate person measures for both genders. As a group, the items also had good quality. As illustrated in Figure 6, all the person fit values were between

0.50 and 1.50. This result indicated that the students responded to the items as projected by the Rasch model.

The items yielded a relatively low person reliability index (PRIX = .34). The person reliability index implied that there was some degree of uncertainty in the estimations of the student locations on the variable map (Figure 4) and that the test (with 17 items and without the time limit) did not provide invariant measurement of originality skills. Although the order of the students (i.e., student locations on the variable map) would most likely change if the students were given the TTCT-figural Form B under the same conditions as the present study, a low amount of variation is preferable for the TTCT-figural Form A because it is a timed test.

Elaboration

The rating scale model was used to examine the following five assumptions about the activities regarding elaboration: Fit is good, average difficulty is the same for all three activities, the response options are equally distributed, the activities behave the same for each gender, and the activity scores involve the same amount of standard error of measurement, which is at the same level along the latent trait continuum.

Fit

The infit mean square values were estimated to be between 0.80 and 1.20 for all three activities. The infit mean square statistics implied that the variation in the data was not considerably larger or smaller than the rating scale model predicted. The standardized infit values were between -2.00 and 2.00 for all three activities, indicating that a good fit for each of the three activities was in fact highly possible. Results suggested that the scoring method for elaboration yielded the measurement of a single construct and that the elaboration score of an activity was not correlated with the elaboration score of another activity—after the latent trait

(i.e., elaboration) was partialed out. The findings of the study confirmed the assumption that the activities have a good fit regarding elaboration.

Average Difficulty

The activities ranged from -0.83 to 0.86 logits with regard to average difficulty. Activity 1 was the easiest in terms of elaboration, while Activity 2 was the most difficult. The average difficulty level of Activity 3 was -0.04 logits. The strata value was estimated to be 8.32. Note that the value is bigger than the number of activities. This result was obtained due to the polytomous scoring. The strata value implied that all three activities with the 6-point scale currently used to score elaboration captured eight discrete difficulty levels. The sufficiently high item reliability index (IRIX = .95) indicated that the difficulty levels of the activities (i.e., activity locations on the variable map [Figure 7]) as well as the response options were estimated accurately.

If all three activities were at the same average difficulty level, a maximum strata value around 6.00 would have been estimated. However, because the activities were not at the same average difficulty level and some of the difficulty levels measured by an activity overlapped with some of the difficulty levels measured by another activity, eight different difficulty levels were detected. The overlapping difficulty levels can be seen in Figure 8. The study did not confirm the assumption that the activities are equally difficult regarding elaboration.

Scaling

The Rasch-Andrich thresholds for Activity 1 were -4.00, -1.97, -0.53, -0.69, and 1.05 logit points from the first threshold to the last. The Rasch-Andrich threshold values estimated for Activity 2 were -2.31, -0.28, 2.16, 2.00, and 2.74 logits. Finally, the Rasch-Andrich thresholds estimated for Activity 3 were -3.21, -1.18, 1.26, 1.10, and 1.84 logit points. As can be seen from

these findings, disordered Rasch-Andrich thresholds were obtained in the study. Note that the value for the fourth threshold was lower than the value for the third threshold.

Traditionally, the Rasch-Andrich thresholds are used to determine whether the response options of a polytomously scored item are distributed evenly. However, when disordered Rasch-Andrich thresholds are obtained, the Rasch-Thurstone thresholds are recommended to examine whether the response options of an item are based on equal intervals (Linacre, 2010). This is because the Rasch-Thurstone thresholds are always ordered even if the Rasch-Andrich thresholds exhibit disordering (Linacre, 2010). However, the Rasch-Thurstone thresholds can only be used when the average measure of a lower response option is lower than the average measure of a higher response option (Linacre, 2010).

The average measures of the response options for Activity 1 were as follows: -4.46, -3.53, -1.83, -0.69, 0.02, and 0.90. As can be seen, the average measure of a particular response option was smaller than the average measure of the next response option (The average measures of the response options for Activity 2 and Activity 3 also increased the same way). Therefore, the Rasch-Thurstone thresholds can be utilized to test whether the response options of each activity were distributed equally.

The Rasch-Thurstone thresholds for Activity 1 were (≈ -5.13), -4.11, -1.95, -0.06, 0.53, 1.42, and (≈ 2.27) logit points (The value in the first parenthesis is not a threshold, but it indicates the lowest trait level measured by the activity [i.e., where the first response option begins]. Similarly, the value in the second parenthesis is not a threshold, but it indicates the highest trait level measured by the activity [i.e., where the last response option ends]). The Rasch-Thurstone threshold values of (≈ -3.44), -2.42, -0.26, 1.63, 2.22, 3.11, and (≈ 3.96) were

obtained for Activity 2. The values of (≈ -4.34), -3.32, -1.16, 0.73, 1.32, 2.21, and (≈ 3.06) were the Rasch-Thurstone thresholds for Activity 3.

The Rasch-Thrustone thresholds showed that the response options were not distributed evenly on the latent trait continuum. The distance between any two subsequent response options was not the same across the response options (see Figure 8). For instance, the distance between the second threshold and the third threshold was more than three times the distance between the third threshold and the fourth threshold. In other words, a student who responded to the fifth response option instead of the fourth response option (i.e., exceeded the fourth threshold) needed at least one third of the trait level that was required to respond to the fourth response option instead of the third response option (i.e., to exceed the third threshold). Findings of the study did not confirm the assumption that the response options of the activities are based on equal intervals.

Before moving on to the assumption about differential item functioning, why disordered thresholds were obtained in the present study should be clarified because it was an unexpected finding. According to Rasch measurement theory, an ideal polytomously scored item consists of ordered response options as well as ordered (Rasch-Andrich) thresholds so that a lower-numbered threshold always corresponds to a lower threshold value. It is, however, sometimes observed that the value of a lower-numbered threshold is larger than the value of a higher-numbered threshold—as observed in the present study.

According to Linacre (2010), threshold disordering does not occur because the lower response option is more difficult than the higher response option (i.e., the lower response option corresponds to a higher trait level). Apart from that, disordered thresholds are observed when a few examinees respond to the intermediate response options (Linacre, 2010). As long as the

average measures of the response options increase hierarchically and the outfit mean square value for each response option is within the suggested range, 0.70 and 1.30, disordering thresholds can be neglected (Linacre, 1999).

Previously, researchers were advised to collapse the response options that corresponded to the disordered thresholds (Linacre, 2002b). However, Linacre (2010) argued that disordered thresholds were indications of transitional categories and could yield useful interpretations. A disordered threshold suggests that a particular response option measures a narrow section of the variable; thus, a low number of respondents are observed in this particular category (Linacre, 2010).

As mentioned previously, the average measures of the response options increased hierarchically from the first response option to the last for all three activities. Additionally, the outfit mean square values were within the suggested range for all the response options for all three activities—the outfit mean square values were 1.10, 1.00, 0.70, 0.80, 0.80, and 1.20 respectively. These findings confirmed that a lower response option always corresponded to a lower trait level and indicated that the fourth response option was a transitional category in all three activities.

Differential Item Functioning

The rating scale model provided different average difficulty levels for boys and girls. Disparities in average (activity) difficulty ranged from 0.19 to 0.31 logits. Welch's *t*-test results showed that none of the differences in average difficulty was statistically significant at the $p < .05$ level. Findings indicated that the activities did not favor students of one gender over students of the opposite gender. The study provided supporting evidence for the assumption about differential item functioning.

Information

The information function of each activity was estimated to have a wide peak (see Figure 9). This finding was expected because under Rasch measurement theory, polytomously scored items have wide information functions. The height of the peak increased towards the right end of the information function for each activity. This occurred because the last four response options were clustered at the right side of the information function and delivered more information than the first two response options did in total. The information function of a particular activity decreased towards the far ends of the latent trait continuum. The activities offered the same amount of information on elaboration (i.e., involved the same amount of standard error of measurement), but due to the variation in the average difficulty levels of the activities, each activity provided this information within a different logit range on the latent trait continuum.

The study revealed that the activities provided the same amount of information on elaboration and that the amount of information an activity delivered on elaboration (i.e., the amount of standard error of measurement involved in an activity) was not at the same level along the latent trait continuum. Consequently, the assumption of the standard error of measurement involved in elaboration scores being the same for all elaboration scores was not confirmed. The test information function estimated for elaboration supported this inference. As seen in Figure 44, high scores provided more information on elaboration, meaning that they were more reliable than average and low elaboration scores. Because elaboration scores in the present study were obtained following the standard testing procedures, one may generalize this conclusion for the elaboration scores of second grade students.

Interim Conclusion

Five assumptions about the activities regarding elaboration were tested. Findings of the study provided supporting evidence for the assumptions about fit and differential item (i.e., activity) functioning. The assumptions about average difficulty, scaling, and standard error of measurement were not supported. The item (i.e., activity) reliability index (IRIX = .97) was sufficiently high, and it suggested that invariant calibration of the average difficulty levels of the activities and invariant calibration of the difficulty levels of the response options were achieved. In other words, the locations of the activities and the response options on the variable map (Figure 7) will not change if another group of second graders are administered the test.

Overall, the results indicated that each individual activity possessed sufficient quality for measuring elaboration and providing appropriate person measures for both genders. The person fit values implied that as a group, the activities worked as predicted by the rating scale model for the majority of the students—person fit measures were not estimated for approximately 50 students because the number of details they added in an activity corresponded to the first response option in all three activities. As illustrated in Figure 10, there were a considerably low number of students with infit mean square values above 1.50. The majority of the students had person fit values below 1.50.

There were many students with person fit values below 0.50, but this was not an important issue. Person fit values below 0.50 implied that the activities worked almost perfectly for those students. In other words, those students responded to the response options to which the rating scale model projected them to respond. Person fit values above 1.50, on the other hand, suggested that some of the students behaved unpredictably. In other words, some students scored high in an activity but were predicted to score low (and *vice-versa*). These findings were

obtained most likely because only three activities (i.e., three items) were analyzed in the study to estimate the trait levels of the students.

The relatively low value for the person reliability index (PRIX = .39) implied that there was some uncertainty in the estimations of the student locations on the variable map (Figure 7) and that the test (with 3 activities and the 6-point scale currently used) did not provide invariant measurement of the elaboration skill levels of the students. Because the TTCT-figural Form A is a timed test, low amount of variation in average item (or activity) difficulty is preferable for the polytomously scored variables.

However, high amount of variation in the difficulty levels of the response options of the items (or activities) is desirable regarding the polytomously scored variables. This is because the polytomously scored variables were developed to capture various trait levels in one item. Therefore, the value for the person reliability index should have been much higher for elaboration. The person reliability index suggested that the scaling of the activities did not work efficiently for the second grade students. Results showed that the first three response options captured considerably wide sections of the variable and that few students responded to the last three response options.

Elaboration-I

There are four assumptions currently made about the items in Activity 2 and Activity 3 regarding item-level elaboration. The items are assumed to have a good fit, to be at the same average item difficulty level, and to behave the same for each gender. A fourth assumption is that every additional detail added to a response requires the same amount of increase in the trait level. In order to investigate these assumptions, the rating scale model was utilized.

Item Fit

The infit mean square values for 15 of the items were between 0.80 and 1.20. Only the second item in Activity 3 showed overfit (I-MNSQ = 0.75). The infit mean square values for 15 good fitting items showed that the variation in the item measures was acceptable for these items. This finding indicated that the responses given to each of these items were affected by one factor and independent of one another. The standardized infit statistics for these 15 items supported this inference. The standardized infit values were between -2.00 and 2.00 for the good fitting items, meaning that a good fit for each item was in fact highly probable.

The infit mean square value for the second item in Activity 3 (I-MNSQ = 0.75) indicated that there was 25% less variation in the item measures than the rating scale model was predicted. This 25% less variation denoted the predictability of the responses given to this particular item. Because the total number of details on this item was higher than those of the other items, the infit mean square value for this particular item implied that one can be sure that any second grade student most likely adds several details to his or her response to this item compared to the other 15 items.

The standardized infit value for the second item in Activity 3 (I-STZD = -2.40) implied that overfit was in fact highly possible. However, 25% less variation can be neglected because Adams and Khoo (1996; as cited in Wilson, 2005, p. 129) argued that up to 25% less variation in the item measures than Rasch measurement theory predicted did not degrade the person measures. Additionally, overfit is not a crucial issue for the TTCT-figural Form A because the items do not give clues about what to do with any other item on the test. It should be noted that a potential scoring error for this item may have caused this overfit—the interrater reliability coefficient was .84 for elaboration-I.

The study showed that the variation in the item measures was acceptable for 15 of the items. The second item in Activity 3 was estimated to involve 25% less variation in the item measures than the rating scale model predicted, but this variation was negligible. Findings of the study confirmed the assumption about item fit for 15 items and implied that the second item in Activity 3 had an acceptable fit.

Average Item Difficulty

The items were clustered between -0.22 and 0.25 logit points with respect to average item difficulty. Both Activity 2 and Activity 3 were composed of items with average difficulty values either below or above the 0 logit point. There was a 0.23-logit difference between the easiest item and the most difficult item in Activity 3, but this disparity can be considered negligible. The strata value of 2.70 was estimated. The strata value implied that two difficulty levels were measured by the items as a group.

As seen in Figure 12, each intermediate response option measured a very narrow portion of the variable, and all the intermediate response options were clustered in the middle. Therefore, each item behaved like a 3-point scaled item, which involved two different difficulty levels. Because the items measured two difficulty levels as a group and each individual item involved two difficulty levels, one can conclude that there was one group of items in the TTCT-figural Form A regarding elaboration-I. In other words, the items in Activity 2 and Activity 3 were not different from each other with respect to average item difficulty. The item reliability index (IRIX = .76) denoted a considerably high probability that the difficulty levels of the items (i.e., item locations on the variable map [Figure 11]) as well as the response options were estimated accurately. Findings of the study implied that the assumption of the items being equally difficult regarding elaboration-I was possibly true.

Item Scaling

For a hypothetical item with the average item difficulty of 0, the Rasch-Andrich thresholds were -0.14, -0.20, -0.03, 0.36, 0.41, and -0.40 logit points. As seen, disordered Rasch-Andrich thresholds were obtained in the study. The value for the second threshold was lower than the value for the first threshold, and the value for the sixth threshold was lower than the values of all the other thresholds. Because the Rasch-Andrich thresholds were disordered, the Rasch-Thurstone thresholds might need to be considered.

The average measures and the outfit mean square values for the response options indicated that the Rasch-Thurstone thresholds could be evaluated to investigate whether every additional detail added to a response required the same amount of increase in the trait level. The average measures of a hypothetical item ($\delta = 0.00$) were -0.90, -.65, -0.45, -0.37, -0.22, -0.12, and 0.11 respectively. The outfit mean square values for the same hypothetical item were 1.00, 1.10, 0.90, 0.90, 0.90, 1.30, and 1.00 in sequence.

For a hypothetical item with the average item difficulty of 0, the Rasch-Thurstone thresholds were ($\approx$ -1.60), -0.80, -0.33, -0.05, 0.19, 0.39, 0.65, and ($\approx$ 1.25) logit points (The value in the first parenthesis is not a threshold, but it indicates the lowest trait level measured by the item [i.e., where the first response option begins]. Similarly, the value in the second parenthesis is not a threshold, but it indicates the highest trait level measured by the item [i.e., where the last response option ends]).

These findings showed that the distance between any two subsequent response options was not the same across the response options for this hypothetical item. For instance, the distance between the first threshold and the second threshold was almost twice the distance between the second threshold and the third threshold. In other words, a student who added 3

details to response instead of 2 details (i.e., exceeded the third threshold) needed almost a half of the trait level that was required to add 2 details to the response instead of 1 detail (i.e., to exceed the second threshold). These inferences apply to the actual test items as well because all the test items have the same structure as the hypothetical item (see Figure 12).

Findings of the study revealed that the same trait level was required to add the same number of details to a response to across the items because all the items were essentially at the same average difficulty level. Another finding was that every additional detail added to a response did not require the same amount of increase in the trait level. Consequently, the assumption about the items regarding elaboration-I (i.e., every additional detail added to a response requires the same amount of increase in the trait level) was not confirmed.

Differential Item Functioning

Gender-based item analyses estimated different average item difficulty values for boys and girls. Average item difficulty differences were between 0.01 and 0.19 logits. Welch's *t*-test results revealed that none of the differences in average item difficulty was statistically significant at the $p < .05$ level. This finding indicated that the items were not in favor of students of one gender over students of the opposite gender. The study confirmed the assumption that the items in the TTCT-figural Form A do not behave differently for boys and girls.

Interim Conclusion

Four assumptions about the items regarding item-level elaboration were investigated in the study. Findings supported the assumption about item fit, average item difficulty, and differential item functioning. Analyses showed that every additional detail added to a response required a different amount of increase in the trait level, so this assumption was not confirmed. Although there was considerable variation in the elaboration-I skill levels of the students, the

variation was not high enough to provide invariant item calibration. The item reliability index was .76. Thus, there is a chance that the locations of some of the items on the variable map (Figure 11) will change if another group of second graders are administered the TTCT-figural Form A.

Overall, the results indicated that each individual test item possessed sufficient quality for measuring elaboration-I and providing appropriate person measures for both genders. As a group, the items worked effectively. There were a considerably low number of students with the person fit values above 1.50 or below 0.50 (see Figure 13).

The items were in the same group regarding average item difficulty, but there was acceptable variation in each individual item due to the polytomous scoring. Therefore, a considerably high value was estimated for the person reliability index (PRIX = .75). This value denoted a high percentage of probability that the estimations of the student locations on the variable map (Figure 11) were accurate. The majority of the student locations on the variable map would remain the same if they were given the TTCT-figural Form B under the same conditions as the present study.

Abstractness of Titles

The following five assumptions are currently made about the item in Activity 1 and the items in Activity 2 regarding abstractness of titles: The items have a good fit, the items are at the same average difficulty level, each item is composed of equally distributed response options, the items behave the same for both genders, and the items involve the same amount of standard error of measurement, which is at the same level along the latent trait continuum. The rating scale model was utilized to test these assumptions.

Item Fit

The majority of the items were estimated to have infit mean square values between 0.80 and 1.20. The following two items showed misfit: the ninth item (I-MNSQ = 0.79) and the tenth item (I-MNSQ = 1.31) in Activity 2. The infit mean square statistics for nine good fitting items showed that the variation in the item measures was acceptable for each of these items. Findings implied that a single factor affected the responses given to these items regarding abstractness of titles and that the abstractness level of a title given to an item cannot be predicted by the abstractness level of a title given to another item—after the latent trait was controlled. These inferences were supported by the standardized infit statistics. The standardized infit values were between -2.00 and 2.00 for these nine items. The standardized infit statistics suggested that a good fit was highly possible for each good fitting item.

According to the conservative approach on infit mean square statistics (Bond & Fox, 2001; Engelhard, 2013), the infit mean square value for the ninth item (I-MNSQ = 0.79) indicated that there was 21% less variation in the item measures than the rating scale model predicted. The infit mean square value for this particular item denoted some degree of predictability of the student responses given to this item. However, the standardized infit value for the same item (I-STZD = -1.50) suggested that misfit was not likely. Additionally, 21% less variation in the item measures was negligible according to Adams and Khoo (1996; as cited in Wilson, 2005, p. 129).

The infit mean square value for the tenth item in Activity 2 (I-MNSQ = 1.31) indicated that there was 31% more variation in the item measures than the rating scale model predicted. This variation indicated randomness in the responses given to this item and suggested that a second factor, other than abstractness of titles, might have impacted the student responses. The

standardized infit value for this item (I-STZD = 1.70), however, suggested that the existence of a second factor was not likely. In addition to this, 31% more variation was not large enough to confirm the existence of a second factor according to Adams and Khoo (1996; as cited in Wilson, 2005, p. 129).

The study showed that the variation in the item measures was acceptable for nine of the items. The ninth item in Activity 2 was estimated to involve 21% less variation and the tenth item in the same activity involved 31% more variation than the rating scale model projected. However, the variation in the item measures was not detrimental for either of these two items. Findings of the study confirmed the assumption about item fit for nine items and implied that the ninth item and the tenth item in Activity 2 had an acceptable fit.

Average Item Difficulty

The lower logit point of -1.89 and the upper logit point of 0.41 were estimated for average item difficulty. The majority of the items were located between -0.21 and 0.41 logit points. The item in Activity 1 was estimated to be the easiest in terms of abstractness of titles. The overwhelming majority of the items in Activity 2 were above the 0 logit point. Only the difficulty level of the ninth item was below the 0 logit point.

The strata value was estimated to be 5.64. The strata value implied that the items measured five different difficulty levels as a group. Note that the number of the difficulty levels implied by the strata value was larger than the number of the response options. This finding suggested that the items were not at the same difficulty level.

In fact, Figure 15 shows that the item in Activity 1 was not at the same average difficulty level as the items in Activity 2. As seen in Figure 15, all the items in Activity 2 captured the difficulty levels between -3.20 and 5.00 logits, while the item in Activity 1 measured the

difficulty levels between -5.00 and 3.00 logits. This figure implied that three of the difficulty levels measured by the item in Activity 1 overlapped with three of the difficulty levels measured by the items in Activity 2. The item in Activity 1 measured one other difficulty level which was not measured by the items in Activity 2. Similarly, the items in Activity 2 measured one other difficulty level which was not measured by the item in Activity 1.

The item reliability index (IRIX = .94) indicated that the difficulty levels of the items (i.e., item locations on the variable map [Figure 14]) as well as the response options were estimated accurately. Findings of the study did not support the assumption that the item in Activity 1 and all 10 items in Activity 2 are equally difficult with respect to abstractness of titles. It should be noted, however, that all 10 items in Activity 2 were at the same difficulty level.

Item Scaling

The Rasch-Andrich threshold values for a hypothetical item with the average item difficulty of 0 were -1.74, -1.90, and 3.63 logits from the first threshold to the last. As can be seen from the findings, disordered Rasch-Andrich thresholds were obtained in the study. The value for the second threshold was lower than the value for the first threshold. Due to the disordered Rasch-Andrich thresholds, the Rasch-Thurstone thresholds should be considered to investigate whether the response options of the items were based on equal intervals. The average measures of the response options suggested that the Rasch-Thurstone thresholds can be considered—the average measures of the response options for a hypothetical item ($\delta = 0.00$) were -3.63, -2.54, -1.25, and -0.63 respectively.

The Rasch-Thrustone threshold values for a hypothetical item with the average item difficulty of 0 were ($\approx$ -3.15), -2.26, -1.37, 3.62, and ($\approx$ 4.70) logit points from the first threshold to the last (The value in the first parenthesis is not a threshold, but it indicates the lowest trait

level measured by the item [i.e., where the first response option begins]. Similarly, the value in the second parenthesis is not a threshold, but it indicates the highest trait level measured by the item [i.e., where the last response option ends]).

As seen, the distance between any two subsequent response options was not the same across the response options for this hypothetical item. For instance, the distance between the first threshold and the second threshold was less than one fifth of the distance between the second threshold and the third threshold. In other words, a student who responded to the fourth response option instead of the third response option (i.e., exceeded the third threshold) needed more than five times the trait level that was required to respond to the third response option instead of the second response option (i.e., to exceed the second threshold). These inferences apply to the actual test items as well because all the test items have the same structure as the hypothetical item (see Figure 15). Findings of the study did not support the assumption that the response options were based on equal intervals.

One point regarding the fourth response option needs to be explained. The outfit mean square value for the fourth response option was 1.60, while the outfit mean square values were within the suggested range for the first three response options—the outfit mean square values for the first three response options were 1.10, 1.00, and 0.90 respectively. The outfit mean square value for the fourth response option indicated that the fourth response option (i.e., an abstract title) appeared randomly in the study. In other words, the students who gave abstract titles to figural responses (i.e., responded to the fourth response option) were not predicted to give abstract titles because they did not respond to the higher response options in the other items. Results implied that second graders rarely give abstract titles to figural responses. This is most likely because second grade students' abstract thinking skills are not fully developed.

Differential Item Functioning

The rating scale model estimated unequal average item difficulty levels for each gender. Differences in average item difficulty ranged from 0.05 to 0.98 logits. According to Welch's *t*-test results, the disparity in average item difficulty was statistically significant for the second item ($t(113) = -2.73$, $p = .01$; easier for girls) and the ninth item ($t(129) = 2.32$, $p = .02$; easier for boys) in Activity 2. Results showed that two items did not behave the same for the boys and girls in the study. Findings of the study indicated that the assumption about differential item functioning might not be true for the second item as well as the ninth item in Activity 2. It should be noted that a potential scoring error may have led to the estimation of differential item functioning for these two items—the interrater reliability coefficient was .89 for abstractness of titles.

Item Information

Because the items were scored in a polytomous manner, the information function of a particular item had two peaks (see Figure 16). One peak was higher on the right side of the information function, and the other peak was higher on the left side of the information function. However, the peak on the left side was higher because the second response option provided some information as well, although it measured a narrow section of the variable. The information function of an item declined as the function approached the far ends of the latent trait continuum. The test items were estimated to offer the same amount of information on abstractness of titles (i.e., to involve the same amount of standard error of measurement). However, due to the variation in the average difficulty levels of the items, each item provided this information within a different logit range on the latent trait continuum.

The study showed that all the items delivered the same amount of information on abstractness of titles and that the amount of information an item provided on abstractness of titles (i.e., the amount of standard error of measurement involved in an item) was not at the same level along the latent trait continuum. Findings indicated that the standard error of measurement involved in scores on abstractness of titles was not the same for all abstractness of titles scores. This conclusion was supported by the test information function estimated with 11 items. The test information function showed that low abstractness of titles scores were more reliable than average and high scores on abstractness of titles (see Figure 44). This finding indicated that when the test is administered under the standard testing conditions, some abstractness of titles scores will be more reliable than others.

Interim Conclusion

Findings of the present study supported only the assumption about item fit. The assumptions about average item difficulty, item scaling, differential item functioning, and item information did not receive statistical support—it should be noted that the items in Activity 2 were at the same difficulty level and that nine items behaved the same for each gender. The sufficiently high item reliability index estimated in the study (IRIX = .94) implied that invariant calibration of the average difficulty levels of the items and invariant calibration of the difficulty levels of the response options were obtained. In other words, the locations of the items on the variable map (Figure 14) will not change if another group of second graders are administered the test.

Overall, results indicated that each individual test item possessed sufficient quality for measuring abstractness of titles and that the majority of the items could provide appropriate person measures for both genders. The items worked effectively as a group. The person fit

statistics showed that the overwhelming majority of the person fit values were below 1.50 (see Figure 17). Person fit values above 1.50 implied that the abstractness levels of the titles given by some students varied randomly. In other words, those students sometimes responded to a lower response option (e.g., gave a single class title) when they were predicted to respond to a higher response option (and *vice-versa*). There were several students with person fit values under 0.50. Person fit values below 0.50 indicated that a considerable number of the students responded to the exact response options that were projected by the rating scale model.

Because the items were scored in a polytomous manner, a moderately high person reliability index (PRIX = .56) was estimated. This value indicated that the student locations on the variable map (Figure 14) will most likely change and that the test (with 11 items and without the time limit) did not provide invariant measurement of the abstractness of titles skill levels of the students. Because each individual item is intended to capture various trait levels through the polytomous scoring, the person reliability index should have been higher. Findings implied that the scoring method for abstractness of titles did not work efficiently for the second grade students in the study. However, this is most likely because the abstract thinking skills of second grade students are not fully developed.

Resistance to Premature Closure

Regarding resistance to premature closure, the following five assumptions are currently made about the items in Activity 2: The items have a good fit, the items are at the same average difficulty level, each item is comprised of evenly distributed response options, the items behaved the same for each gender, and the items involve the same amount of standard error of measurement, which is at the same level along the latent trait continuum. These assumptions were tested using the rating scale model.

Item Fit

The infit mean square values were estimated to be between 0.80 and 1.20 for all the items in Activity 2. Findings showed that the variation in the item measures was within the recommended range (Bond & Fox, 2001; Engelhard, 2013), indicating that a single factor affected the responses given to the items regarding resistance to premature closure. Another indication was that the response given to an item did not give the examinee a hint about what to do with any other incomplete figure in Activity 2. The standardized infit values confirmed that a good fit was most likely for each item in the activity. All the standardized infit values were between -2.00 and 2.00. The present study provided supporting evidence for the assumption that the items have a good fit regarding resistance to premature closure.

Average Item Difficulty

The items were located between -0.73 and 0.71 logit points in terms of average item difficulty; all the items were between -0.73 and 0.71 logits. The strata value was estimated to be 5.53. The strata value suggested that the items measured five different difficulty levels as a group regarding resistance to premature closure. Note that the number of the difficulty levels implied by the strata value was larger than the number of the response options. This finding suggested that the items were not at the same difficulty level.

Figure 19 shows that the first, second, third, seventh, eighth, ninth, and tenth items were at the same average difficulty level. Figure 19 also shows that the fourth and sixth items were at the same difficulty level, while the fifth item was at another difficulty level. Figure 19 indicates that two of the difficulty levels measured by the fourth and sixth items overlapped with the two of the difficulty levels measured by the first, second, third, seventh, eighth, ninth, and tenth items. Similarly, two of the difficulty levels measured by the fifth item overlapped with the two

of the difficulty levels measured by the first, second, third, seventh, eighth, ninth, and tenth items. The fourth and sixth items measured one other difficulty level between -2.00 and -1.00 logits, while the fifth item measured another difficulty level between 1.00 and 2.00 logits—these two additional difficulty levels were not measured by the rest of the items.

The item reliability index (IRIX = .94) indicated that the difficulty levels of the items (i.e., item locations on the variable map [Figure 18]) as well as the response options were estimated accurately. Findings of the study did not support the assumption that the items are equally difficult regarding resistance to premature closure.

Item Scaling

The Rasch-Andrich thresholds were 0.46 and -0.46 logit points in sequence for a hypothetical item with the average item difficulty of 0. As seen, disordered Rasch-Andrich thresholds were obtained in the study. The value for the second threshold was lower than the value for the first threshold. Because the Rasch-Andrich thresholds were disordered, the Rasch-Thurstone thresholds should be considered to investigate whether the response options were based on equal intervals. The average measures of the response options increased hierarchically—the average measures were -0.44, 0.05, and 0.54 in sequence—and the outfit mean square values were 1.00, 1.00, and 1.10 respectively. These findings showed that the Rasch-Thurstone thresholds could be utilized to address the assumption about item scaling.

The Rasch-Thurstone thresholds were (≈ -1.00), -0.31, 0.31, and (≈1.00) logit points in sequence for a hypothetical item with the average item difficulty of 0 (The value in the first parenthesis is not a threshold, but it indicates the lowest trait level measured by the item [i.e., where the first response option begins]. Similarly, the value in the second parenthesis is not a

threshold, but it indicates the highest trait level measured by the item [i.e., where the last response option ends]).

These findings showed that the distance between the first threshold and the second threshold was not the same as the distance between the first threshold and the lowest range of the item (≈ -1.00 logit point) as well as the distance between the second threshold and the highest range of the item (≈ 1.00 logit point). For instance, a student who responded to the third response option instead of the second response option (i.e., exceeded the second threshold) needed approximately 0.10-logit lower trait level than the trait level that was required to respond to the second response option instead of the first response option (i.e., to exceed the first threshold).

Although the findings showed that the response options were not distributed evenly on the latent trait continuum from a numerical point of view, the distance between any two subsequent response options can be considered to be the same because the 0.10-logit difference was considerably small (see Figure 19). Therefore, one can conclude that the assumption that the response options of the items are distributed equally is possibly true regarding resistance to premature closure.

Differential Item Functioning

Analyses yielded different average item difficulties for each gender. Average item difficulty differences ranged from 0.03 to 0.50 logits. According to Welch's *t*-test results, the disparity in average item difficulty was statistically significant for the seventh item in Activity 2 ($t(152) = 2.41$, $p = .017$). The item was 0.50 logits easier for boys. Findings indicated that the tension created by the shape of the stimulus was significantly lower for boys, so boys had a higher chance of leaving the stimulus open or closing it with irregular lines. It is possible that a

potential scoring error may have led to the estimation of this differential item functioning—the interrater reliability coefficient was .89 for resistance to premature closure.

Item Information

The test items were scored in a polytomous manner for resistance to premature closure. Therefore, the items were expected to have wide information functions. However, because the second response option corresponded to a narrow section of the variable, analyses yielded bell-shaped information functions for the test items (see Figure 20). The midpoint of the information function of a particular item was the item's average difficulty point on the latent trait continuum. The amount of information provided by the item on resistance to premature closure was the highest at this average item difficulty point—this point was where the standard error of measurement was the smallest. The amount of information provided by an item on resistance to premature closure declined in the direction of the far ends of the latent trait continuum. The test items delivered the same amount of information on resistance to premature closure. However, due to the variation in average item difficulty, each item provided this information within a different portion of the latent trait continuum.

The study showed that all the items offered the same amount of information on resistance to premature closure and that the amount of information an item provided on resistance to premature closure (i.e., the amount of standard error of measurement involved in an item) was not at the same level along the latent trait continuum. Findings suggested that the standard error of measurement involved in scores on resistance to premature closure was not the same for all resistance to premature closure scores. The test information function estimated for scores on resistance to premature closure supported this conclusion. The average scores on resistance to premature closure provided the highest amount of information on resistance to premature closure

(see Figure 44). This finding suggested that when the test is administered under the standard testing conditions, some resistance to premature closure scores will be more reliable than others.

Interim Conclusion

Findings of the present study supported the assumptions about item fit and item scaling. The assumptions about average item difficulty, differential item functioning, and item information were not supported—it should be noted that nine of the items did not exhibit differential item functioning. The item reliability index (IRIX = .94) suggested that invariant item calibration was achieved. Therefore, the locations of the items on the variable map (Figure 18) are expected to remain the same if another group of second graders are administered the test.

Overall, the results indicated that each individual test item possessed sufficient quality for measuring resistance to premature closure and that the overwhelming majority of the items could provide appropriate person measures for both genders. The items also worked effectively as a group. A considerably low number of person fit values were above 1.50 (see Figure 21). Person fit values above 1.50 indicated that some students did not respond to the response options that were predicted by the rating scale model. There were a few students with person with values below 0.50. Person fit values below 0.50 indicated that the student responses were predicted perfectly by the rating scale model for those students.

A moderately high person reliability index (PRIX = .65) was estimated in the study. Therefore, some of the student locations on the variable map (Figure 18) are expected to change if the students were given the TTCT-figural Form B. The test (with 10 items and without the time limit) did not provide invariant measurement of the resistance to premature closure skill levels of the students. The variation in average item difficulty and the difficulty levels of

response options was relatively low. Because each individual item is intended to capture various trait levels through the polytomous scoring, the person reliability index should have been higher.

Emotional Expressiveness

Three assumptions are currently made about the test items regarding emotional expressiveness: The items have a good fit, the items are equally difficult, and the items behave the same for boys and girls. The dichotomous Rasch model was used to investigate these assumptions.

Item Fit

The infit mean square values were between 0.80 and 1.20 for all the items with the exception of the following items: the fifth (I-MNSQ = 1.21), sixth (I-MNSQ = 0.77) and seventh (I-MNSQ = 0.73) items in Activity 2. The infit mean square values for 14 good fitting items implied that a single factor affected the responses given to the items regarding emotional expressiveness and that the response patterns were not too deterministic. These conclusions were supported by the standardized infit statistics. The standardized infit values were between -2.00 and 2.00 for these 14 items, suggesting that a good fit was highly possible for each item.

The infit mean square value for the fifth item in Activity 2 (I-MNSQ = 1.21) indicated that there was 21% more variation in the item measures than the Rasch model predicted. This variation denoted some degree of randomness in the responses given to this item, but 21% more variation in the item measures was negligible according to Adams and Khoo (1996; as cited in Wilson, 2005, p. 129). Additionally, the standardized infit value for this particular item (I-STZD = 1.50) implied that misfit was not likely.

The infit mean square value for the sixth item in Activity 2 (I-MNSQ = 0.77) showed that the variation in the item measures was 23% lower than the Rasch model predicted. The variation

in the item measures indicated that the responses given to this item were too deterministic—it should be noted that only nine students expressed emotion in this item. However, the standardized infit value for this item (I-STZD = -0.80) suggested that misfit was not likely. Additionally, according to Adams and Khoo's (1996; as cited in Wilson, 2005, p. 129) argument, 23% less variation in the item measures for this item was not detrimental.

Item analyses showed that the responses given to the seventh item in Activity 2 were also too deterministic—only four students expressed emotion in this item. The infit mean square value for this item (I-MNSQ = 0.73) denoted 27% less variation in the item measures than the Rasch model projected. The amount of variation in the item measures was 2% lower than the acceptable amount of variation recommended by Adams and Khoo (1996; as cited in Wilson, 2005, p. 129). However, 27% less variation in the item measures was possibly not harmful because the items in the TTCT-figural Form A do not give clues about what to do with any other item. Additionally, the standardized infit value for this item (I-STZD = 0.50) suggested that overfit was not likely.

The study showed that the variation in the item measures was acceptable for 14 of the items. The fifth, sixth, and seventh items in Activity 2 were estimated to involve more (for the fifth item) or less (for the sixth and seventh items) variation in the item measures than the Rasch model projected. However, the standardized infit values suggested that misfit was not likely for any of these items. These results implied that the variation in the responses given to the fifth, sixth, and seventh items in Activity 2 did not degrade the person measures. Findings of the study supported the assumption about item fit for 14 items and implied that the fifth, sixth, and seventh items in Activity 2 had an acceptable fit.

Item Difficulty

The items were spread between -2.59 and 1.22 logit points with respect to difficulty. The majority of the items were between -0.40 and 1.22 logits. The item in Activity 1 was the easiest in terms of expressing emotion in a response. The items in Activity 2 were either below or above the 0 logit point. Some stimuli in Activity 2 (e.g., the fifth item; $\delta = -1.73$) generally induced emotion in a response, while some other stimuli (e.g., the second item; $\delta = .92$) did not lead the students to responses with emotion expressed. The first six items in Activity 3 were all above the 0 logit point with respect to item difficulty. There was a 1.18-logit difference between the easiest item and the most difficult item in Activity 3, but only five manifestations of emotional expressiveness led to the estimation of this difference.

The strata value of 3.10 was estimated, indicating that there were three discrete item groups on the test with regard to difficulty. The item reliability index was considerably high (IRIX = .81), but it was below the cutoff point suggested for invariant item calibration. In other words, there is a considerably—but not sufficiently—high probability that these item groups existed and that the difficulty levels of the items (i.e., item locations on the variable map [Figure 22]) were determined accurately. Findings of the study did not support the assumption that the items are equally difficult regarding emotional expressiveness.

Differential Item Functioning

The Rasch model estimated different item difficulty levels for each gender. Differences in item difficulty ranged from 0.16 to 1.76 logits. Welch's *t*-test results showed that the difficulty level disparities were significantly different for two items: the fourth item ($t(66) = 2.02$, $p = .047$) and the fifth item ($t(53) = 2.22$, $p = .031$) in Activity 2. The fourth item was easier for boys, while the fifth item was easier for girls. Results implied that two of the items on the test did not

behave the same for boys and girls. Findings of the study suggested the assumption about differential item functioning might not be true for two items. It should be noted that a possible scoring error for these two items may have caused the differential item functioning, although the interrater reliability coefficient for emotional expressiveness was considerably high (r = .89).

Interim Conclusion

Three assumptions about the items regarding emotional expressiveness were investigated in the present study. The assumption about item fit received statistical support. The assumptions about item difficulty and differential item functioning were not supported by the findings. Analyses of differential item functioning implied that two items might not behave the same for each gender. The considerably high item reliability index (IRIX = .81) denoted a high chance of invariant item calibration; however, the item reliability index indicated that the locations of some of the items on the variable map (Figure 22) may change if another group of second graders are administered the test.

Overall, the results indicated that each test item possessed sufficient quality for measuring emotional expressiveness and that the overwhelming majority of the items could provide appropriate person measures for both genders. The items work efficiently as a group for a considerable number of the students—person fit values were not estimated for approximately 100 students because they did not manifest emotional expressiveness in any item. There were a few students with person fit values above 1.50 or below 0.50 (see Figure 23). The overwhelming majority of the person fit values were between 0.50 and 1.50.

A value of zero was estimated for the person reliability index, implying that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 22) and that the test (with 17 items and without the time limit) did not provide invariant

measurement of the emotional expressiveness skill levels of the students. Therefore, the trait levels of the students would be estimated to be different if they were given the TTCT-figural Form B under the same conditions. No variation is, however, desirable for the TTCT-figural Form A due to its timed nature.

Storytelling Articulateness

The Rasch model was used to examine the following three assumptions about the items regarding storytelling articulateness: The items have a good fit, the items are equally difficult, and the items behave the same for boys and girls.

Item Fit

The infit mean square values were within the recommended range for a good fit for all the items with the exception of the item in Activity 1 (I-MNSQ = 1.22) and the third item in Activity 3 (I-MNSQ = 0.71). The infit mean square statistics for 15 good fitting items suggested that unidimensionality and local independence were not violated for any of these items. This inference was supported by the standardized infit statistics. For all these 15 items, the standardized infit values were between -2.00 and 2.00, and they indicated that a good fit was highly possible for each of good fitting item.

The infit mean square value for the item in Activity 1 (I-MNSQ = 1.22) implied that there was 22% more variation in the item measures than the Rasch model predicted. This variation indicated some degree of randomness in the responses given to this particular item. However, 22% more variation in the item measures was not detrimental according to Adams and Khoo (1996; as cited in Wilson, 2005, p. 129). In addition to this, the standardized infit value for this item (I-STZD = 1.90) implied that misfit was not likely.

The infit mean square value for the third item in Activity 3 (I-MNSQ = 0.71) showed that there was 29% less variation in the item measures than the Rasch model projected. The infit mean square value implied that the responses given to this item were too deterministic—overall, 15 students manifested storytelling articulateness in this item. The amount of variation in the item measures was 4% lower than the acceptable amount of variation recommended by Adams and Khoo (1996; as cited in Wilson, 2005, p. 129). However, 29% less variation in the item measures was probably not harmful because the items on the TTCT-figural Form A do not give clues about what to do with any other item. Additionally, the standardized infit value for this item (I-STZD = -1.60) suggested that overfit was not likely.

The study showed that the variation in the item measures was acceptable for 15 of the items. The item in Activity 1 was estimated to involve 22% more variation and the third item in Activity 3 involved 29% less variation than the Rasch model predicted. However, the standardized infit values suggested that misfit was not likely for either of these items. Findings of the study provided supporting evidence for the assumption about item fit for 15 of the items and indicated that the item in Activity 1 as well as the third item in Activity 3 had an acceptable fit.

Item Difficulty

The difficulty levels of the items ranged from -3.95 to 2.16 logits, with the majority of the items being between -0.79 and 0.75 logits. The item in Activity 1 was the easiest in terms of creating a response with a story. The items in Activity 2 were either below or above 0 logit point; however, the majority of the items were above the 0 logit point. Additionally, the most difficult item on the test was in Activity 2 (the tenth item). Some stimuli in Activity 2 (e.g., the third item; $\delta = -0.79$) generally led students to depict a story in a response, while some other stimuli (e.g.,

the tenth item; $\delta = 2.16$) did not lead the students to responses with a story. Similarly, the first six items in Activity 3 were either below or above the average difficulty level, 0. There was a 0.99-logit difference between the easiest item and the most difficult item in Activity 3—eight manifestations of storytelling articulateness led to the estimation of this difference.

The strata value was estimated to be 4.22. This value implied that there were four distinct item groups on the test in terms of item difficulty. The considerably high item reliability index estimated in the study (IRIX = .89) implied that these item groups were most likely present and that the difficulty levels of the items (i.e., the item locations on the variable map [Figure 24]) were possibly determined accurately. Item difficulty analyses did not support the assumption of the items being equally difficult regarding storytelling articulateness.

Differential Item Functioning

Item analyses estimated unequal item difficulties for each gender. Item difficulty differences ranged from 0.02 to 1.74 logits. The disparity in item difficulty was estimated to be statistically significant for the seventh item in Activity 2 ($t(98) = 2.27$, $p = .025$). The item was 1.74 logits easier for boys. Findings of the study suggested that one item might behave differently for boys and girls in second grade. It should be noted, however, that the differential item functioning detected for this item might be due to the low frequency of the expressions of storytelling articulateness (only 11 expressions were observed on the seventh item) and/or a possible scoring error (although the interrater reliability coefficient was .90).

Interim Conclusion

Three assumptions about the items regarding storytelling articulateness were examined in the present study. The assumption that the items have a good fit received statistical support. The assumption about differential item functioning was confirmed for 16 of the items. Findings

implied that one item might behave differently for each gender. The assumption about item difficulty was not confirmed. The item reliability index (IRIX = .89) denoted a considerably high chance of invariant item calibration. Therefore, the locations of the items on the variable map (Figure 24) will most likely keep their positions if another group of second graders take the test.

Findings of the study suggested that each individual item possessed sufficient quality for measuring storytelling articulateness and that the overwhelming majority of the items could provide appropriate person measures for students of each gender. As a group, the items had acceptable quality for the majority of the students—person fit values were not estimated for approximately 70 students because they did not express storytelling articulateness in any item. The majority of the person fit measures were below 1.50 (see Figure 25). There were several students with person fit values below 0.50, but those students manifested storytelling articulateness only in one item.

The person reliability index was estimated to be zero. The zero value meant that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 24) and that the test (with 17 items and without the time limit) did not provide invariant measurement of the storytelling articulateness skill levels of the students. The zero value for the person reliability index was estimated because the variation in item difficulty was not high enough, although there were four distinct item groups. Low amount of variation in item difficulty is, however, preferable for the TTCT-figural Form A.

Movement or Action

There are three assumptions currently made about the items regarding movement or action. The items are assumed to have a good fit, to be equally difficult, and to behave the same for students of each gender. The Rasch model was utilized to examine these assumptions.

Item Fit

The infit mean square values were between 0.80 and 1.20 for all the items. The infit mean square statistics implied that there was acceptable variation in the item measures regarding movement or action. The standardized infit values were also within the recommended range for a good fit for all the items, and they confirmed that a good fit was in fact highly possible for all the items. These findings suggested that a single factor affected the responses given to the items regarding movement or action and that a response given to an item did not give an idea about what to do with the any other item on the test. Findings of the study provided supporting evidence for the assumption about item fit.

Item Difficulty

The lower logit point of -2.53 and the upper logit point of 1.82 were estimated for item difficulty. The majority of the items were located between -1.26 and 0.49 logit points. The item in Activity 1 was the easiest item in terms of creating a response with an object or a character that moved. The items in Activity 2 were either below or above the 0 logit point. Additionally, the two most difficult items on the test (the fifth and tenth items) were in Activity 2. Some stimuli in Activity 2 (e.g., the first item; $\delta = -1.26$) generally led the students to draw moving characters or objects in a response. Some other stimuli (e.g., the sixth item; $\delta = 1.82$), on the other hand, were relatively challenging in terms of expressing movement or action in a response. The first six items in Activity 3 were all above the average difficulty level, 0. There was a 1.10-logit difference between the easiest item and the most difficult item in Activity 3—eight manifestations of movement or action led to the estimation of this difference.

The strata value of 4.09 was estimated, suggesting that there were four distinct item groups on the test with respect to item difficulty. The item reliability index (IRIX = .89)

indicated a considerably high possibility that these item groups existed and that the difficulty levels of the items (i.e., the item locations on the variable map [Figure 26]) were determined accurately. Findings of the study did not support the assumption about item difficulty regarding movement or action.

Differential Item Functioning

Gender-based item analyses estimated different item difficulty levels for each gender. Differences in item difficulty were between 0.05 and 1.62 logits. The disparity was statistically significant for the ninth item ($t(110) = -2.11$, $p = .037$) in Activity 2. The item was 1.06 logits easier for girls. Findings of the study suggested that one item on the test might not behave the same for boys and girls. It should be noted, however, that the differential item functioning detected for this item might be due to the low frequency of the expressions of movement or action (only 9 expressions were observed on the ninth item) and/or a possible scoring error, although the interrater reliability coefficient was considerably high ($r = .89$).

Interim Conclusion

Three assumptions about the items regarding movement or action were tested in the present study. The assumption about item fit was supported by the findings of the study. The assumption about differential item functioning was confirmed for 16 of the items. Results suggested that one item might behave differently for boys and girls. Finally, the assumption about item difficulty did not receive statistical support. The item reliability index was estimated to be considerably high (IRIX = .89), and it implied that that the locations of the items on the variable map (Figure 26) will possibly not change if another group of second graders are administered the test.

The study indicated that each test item possessed sufficient quality for measuring storytelling articulateness and that the overwhelming majority of the items could provide appropriate person measures for students of each gender. As a group, the items worked effectively for the majority of the students—person fit values were not estimated for approximately 50 students because they had a zero score. Nearly all the person fit values were between 0.50 and 1.50 (see Figure 27).

The person reliability index was estimated to be zero. This result implied that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 26), meaning that the test (with 17 items and without the time limit) did not provide invariant measurement of the movement or action skill levels of the students. The small amount of variation in item difficulty led to the estimation of zero for the person reliability index, although there were four distinct item groups with respect to difficulty. Because the TTCT-figural Form A is a timed test, low amount of variation in item difficulty is desired.

Expressiveness of Titles

The following three assumptions are currently made about the items regarding expressiveness of titles: The items have a good fit, the items are at the same difficulty level, and the items do not favor students of one gender over students of the opposite gender. In order to test these assumptions, the Rasch model was used.

Item Fit

The infit mean square values were within the recommended range for a good fit for all the items except the item in Activity 1 (I-MNSQ = 1.35) and the first item (I-MNSQ = 0.74) as well as the second item (I-MNSQ = 1.31) in Activity 2. The infit mean square statistics for 14 good fitting items showed that the variation in the item measures was acceptable. Findings implied

that a single factor affected the responses given to these items regarding expressiveness of titles and that the response given to an item cannot be predicted by the response given to another item—after the latent trait (i.e., expressiveness of titles) was controlled. These inferences were supported by the standardized infit statistics. The standardized infit values were also within the recommended range for these 14 items. The standardized infit statistics suggested that a good fit was highly possible for each good fitting item.

The infit mean square value for the item in Activity 1 (I-MNSQ = 1.35) showed that there was 35% more variation in the item measures than the Rasch model predicted. The variation in the item measures indicated that the responses were given haphazardly to this item and that a second factor, other than expressiveness of titles, might have impacted the student responses. The standardized infit value for this items (I-STZD = 3.60) suggested that misfit was likely. Because the overwhelming majority of the items on the test were estimated to have a good fit, the potential second factor impacting the responses given to this item was most likely not a construct related to creativity. Apart from that, the second factor may have been related to the testing conditions specific to this item such as giving a 10-minute time limit for one item and presenting one stimulus on one page. There is also a possibility that a scoring error for this item may have been the second factor.

The infit mean square value for the second item in Activity 2 (I-MNSQ = 1.31) denoted 31% more variation than the Rasch model predicted. This variation also denoted the haphazardness in the responses given to this item and implied that a second factor might have impacted the student responses. The standardized infit value for this item (I-STZD = 1.30), however, suggested that the existence of a second factor was not likely. Additionally, this 31% more variation can be neglected because Adams and Khoo (1996; as cited in Wilson, 2005, p.

129) argued that up to 33% more variation in the item measures than Rasch measurement theory predicted was not large enough to confirm the existence of a second factor.

Finally, the infit mean square value for the first item in Activity 2 (I-MNSQ = 0.74) implied that there was 26% less variation in the item measures than the Rasch model projected. The infit mean square statistics for this item denoted the predictability of the responses given to this item. The amount of variation in the item measures was 1% lower than the minimum amount of acceptable variation recommended by Adams and Khoo (1996; as cited in Wilson, 2005, p. 129). However, 26% less variation in the item measures can be neglected because it shows that the item worked more efficiently than the Rasch model predicted. Additionally, the standardized infit value for this item (I-STZD = -1.80) implied that misfit was not likely.

The study showed that the overwhelming majority of the items had a good item fit, meaning that the amount of variation in the item measures was acceptable for these items. The infit mean square and standardized infit values for the item in Activity 1 indicated that the responses were given randomly to this item. In other words, students who manifested expressiveness of titles in their responses given to this item were not predicted to manifest expressiveness of titles (and *vice versa*).

There was more variation in the item measures than the Rasch model projected for the second item in Activity 2. For the first item in the same activity, there was less variation in the item measures than the Rasch model predicted. However, the standardized infit values for these two items suggested that misfit was not likely. Findings of the study implied that 14 items had a good fit and that the first item as well as the second item in Activity 2 had an acceptable fit. Results indicated that a second factor related to the testing conditions or a scoring error specific to the item in Activity 1 might have impacted the student responses given to this item.

Item Difficulty

The items were located between -3.15 and 0.90 logit points with respect to item difficulty. The majority of the items were clustered between -0.35 and 0.90 logit points. The item in Activity 1 was the easiest item in terms of giving an expressive title to a response. The items in Activity 2 were either below or above the 0 logit point. Some stimuli in Activity 2 (e.g., the first item; $\delta = -0.86$) generally led the students to give an expressive title to their responses, while some other stimuli (e.g., the tenth item; $\delta = 0.53$) did not necessarily yield expressive titles. The first six items in Activity 3 were all above the 0 logit point. There was a 0.83-logit difference between the easiest item and the most difficult item in Activity 3—eight manifestations of expressiveness of titles led to the estimation of this difference.

The strata value of 3.85 was estimated, implying that there were three different item groups on the test regarding item difficulty. The considerably high item reliability index estimated in the study (IRIX = .87) implied that these item groups most likely existed and that the difficulty levels of the items (i.e., item locations on the variable map [Figure 28]) were possibly determined with high accuracy. Item difficulty analyses did not support the assumption of the items being equally difficult regarding expressiveness of titles.

Differential Item Functioning

Difficulty levels of the items were different for each gender. Differences in item difficulty ranged from 0.05 to 1.94 logits. Welch's *t*-test results showed that the disparities in item difficulty were statistically significant for the seventh item ($t(94) = 2.04$, $p = .044$; easier for boys) in Activity 2 and the first item ($t(66) = -2.35$, $p = .022$; easier for girls) as well as the third item ($t(85) = 2.72$, $p = .008$; easier for boys) in Activity 3. Findings of the study suggested that three items might not behave the same for each gender. It should be noted that a potential scoring

error may have caused the estimation of differential item functioning for these three items—the interrater reliability coefficient for expressiveness of titles was .90.

Interim Conclusion

Three assumptions about the items regarding expressiveness of titles were investigated in the present study. The assumption about item fit received statistical support for 16 of the items. Both the infit mean square and standardized infit values suggested that a second factor might have impacted the student responses given to the item in Activity 1. Because the other items had a good fit, this second factor was probably related to the testing conditions or a scoring error specific to this item. Findings of differential item functioning suggested that three items might behave differently for boys and girls. Finally, the assumption about item difficulty was not supported. The item reliability index was estimated to be considerably high (IRIX = .87), and it implied that that the locations of the items on the variable map (Figure 28) will most likely keep their relative positions if another group of second graders are administered the test.

Findings of the study indicated that the majority of the test items possessed sufficient quality for measuring expressiveness of titles and providing appropriate person measures for both genders. As a group, the items had acceptable quality. The majority of the person fit measures were below 1.50 (see Figure 29). There were several students with person fit values below 0.50, but those students manifested expressiveness of titles only in one item. It should be noted that approximately 70 students had a zero score. Therefore, person fit measures were not estimated for those students.

A person reliability index of .24 was estimated in the study. This relatively low value implied that there was some degree of uncertainty in the estimations of the student locations on the variable map (Figure 28) and that the test (with 17 items and without the time limit) did not

provide invariant measurement of the expressiveness of titles skill levels of the students. This low value for the person reliability index was estimated because the variation in item difficulty was not high enough, although there were three distinct item groups on the test. However, low variation in item difficulty is preferred for the TTCT-figural Form A.

Synthesis of Incomplete Figures

Torrance (1979) stated in his book that synthesis of incomplete figures was not manifested by the test takers as often as other variables were. Findings of the study supported Torrance's statement. Only two of the students in the present study combined two or more incomplete figure in Activity 2. Therefore, no interpretable results were obtained for the items regarding synthesis of incomplete figures. Because a considerably low number of students expressed synthesis of incomplete figures in the present study, synthesis of incomplete figures needs to be investigated with a larger sample size to determine whether there are particular response patterns exhibited by second grade students or whether second grade students manifest synthesis of incomplete figures in a random manner.

Synthesis of Lines

Torrance (1979) acknowledged that synthesis of lines was not expressed by many test takers, but he stated that synthesis of lines was observed more often than synthesis of incomplete figures. Findings of the study supported Torrance's statement. Six of the students participating in the study combined two or more parallel lines in Activity 3. No interpretable results were obtained for synthesis of lines. Like synthesis of incomplete figures, synthesis of lines was manifested by a considerably low number of students in the study. Synthesis of lines should be examined with a larger sample size to determine whether there are particular response patterns

exhibited by second grade students or whether second grade students combine two of more parallel lines haphazardly.

Unusual Visualization

The following three assumptions are currently made about the items regarding unusual visualization: The items have a good fit, the items are at the same difficulty level, and the items do not favor students of one gender over students of the opposite gender. In order to test these assumptions, the Rasch model was used.

Item Fit

The infit mean square values were between 0.80 and 1.20 for all the items. The infit mean square statistics showed that the variation in the item measures was not considerably larger or smaller than the Rasch model predicted. This finding indicated that unidimensionality and local independence were not violated. The standardized infit statistics supported these inferences. All the standardized infit values were within the recommended range for a good fit, and they suggested that a good fit was likely for all the items. Findings of the study provided supporting evidence for the assumption that the items have a good fit regarding unusual visualization.

Item Difficulty

The items were spread between -2.27 and 1.11 logit points with respect to difficulty. The item in Activity 1 was the easiest item in terms of creating a response with unusual perspective. The items in Activity 2 were either below or above the 0 logit point—the most difficult item on the test was in Activity 2 (the tenth item; $\delta = 1.11$). It was easier for the students to create responses with unusual perspective by using some stimuli (e.g., the third item; $\delta = -1.26$). Some other stimuli (e.g., the first item; $\delta = 0.95$) were relatively more difficult in terms of manifesting unusual visualization. The first six items in Activity 3 were all above the 0 logit point. In

general, it was relatively difficult to create a response with unusual perspective by using a particular item in Activity 3. There was a 0.50-logit difference between the easiest item and the most difficult item in Activity 3, but this difference can be considered negligible.

The strata value was estimated to be 5.86. This value suggested that there were five distinct item groups on the test regarding item difficulty. A sufficiently high item reliability index estimated in the study (IRIX = .94) implied that these item groups existed and that the locations of the items on the variable map (Figure 30) were determined accurately. Item difficulty analyses did not support the assumption that the items are equally difficult regarding unusual visualization.

Differential Item Functioning

The Rasch model estimated unequal item difficulty levels for each gender. Differences in item difficulty ranged from 0 to 0.79 logits. None of the differences in item difficulty was statistically significant at the $p < .05$ level according to Welch's *t*-test results. Results indicated that all the items in the TTCT-figural Form A behaved the same for boys and girls. The study provided supporting evidence for the assumption about differential item functioning.

Interim Conclusion

Three assumptions about the items regarding unusual visualization were examined in the present study. The assumptions about item fit and differential item functioning were confirmed. The assumption about item difficulty, on the other hand, did not receive statistical support. The item reliability index (IRIX = .94) denoted a sufficiently high possibility of invariant item calibration. Therefore, the items will keep their relative locations on the variable map (Figure 30) if another group of second graders are administered the test.

Findings of the study indicated that each the test item possessed sufficient quality for measuring unusual visualization and providing appropriate person measures for both genders. As a group, the items worked as the Rasch model predicted. Nearly all the person fit values were between 0.50 and 1.50 (see Figure 31). This finding implied that the students responded to the items as projected by the Rasch model.

The value estimated for the person reliability index (PRIX = .15) indicated that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 30) and that the test (with 17 items and without the time limit) did not provide invariant measurement of the unusual visualization skill levels of the students. This low value was estimated due to the small amount of variation in the difficulty levels of the items, although there were five different items groups in terms of difficulty. Because the TTCT-figural Form A is a timed test, low amount of variation in item difficulty is preferable.

Internal Visualization

Three assumptions are currently made about the items regarding internal visualization. The items are assumed to have a good fit, to be at the same difficulty level, and to behave the same for each gender. These assumptions were examined using the dichotomous Rasch model.

Item Fit

The infit mean square values were within the recommended range for a good fit for all the items. The infit mean square values indicated that the variation in the item measures was not considerably large or small to affect unidimensionality and item independence. The standardized infit values were between -2.00 and 2.00 for all the items, and they indicated that the inferences made about unidimensionality and item independence were most likely true. Findings of the

study provided supporting evidence for the assumption that the items have a good fit regarding internal visualization.

Item Difficulty

The items ranged from -0.93 to 2.22 logits with respect to item difficulty, with the majority of the items being between -0.93 and 0.51 logits. The item in Activity 1 was the easiest item in terms of creating a response with internal visualization. The items in Activity 2 were either below or above the 0 logit point—the most difficult item on the test was in Activity 2 (the tenth item; $\delta = 2.22$). It was relatively easier for the students to manifest internal visualization in the responses given to some items in Activity 2 (e.g., the fifth item; $\delta = -0.47$). Some other items (e.g., the tenth item; $\delta = 2.22$), on the other hand, did not lead the students to express internal visualization in responses. The first six items in Activity 3 were all below the 0 logit point. It was relatively easy for the students to create a response with internal visualization using a particular item in Activity 3. There was a 0.41-logit difference between the easiest item and the most difficult item in Activity 3, but only two expressions of internal visualization led to the estimation of this difference.

The strata value was estimated to be 2.17. The strata value implied that there were two discrete item groups on the test in terms of item difficulty. The item reliability index estimated in the study (IRIX = .65) denoted a moderate possibility of the presence of these two item groups and the accurate determination of the item locations on variable map (Figure 32). Item difficulty analyses suggested that the assumption that the items are equally difficult regarding internal visualization might not be true.

Differential Item Functioning

Item analyses estimated unequal item difficulties for each gender. Differences in item difficulty were between 0.10 and 2.39 logits. Welch's *t*-test results showed that the disparity was statistically significant for the second item in Activity 2 ($t(77) = 2.18$, $p = .032$). The item was 2.39 logits easier for boys. The study suggested that one item on the test might be in favor boys. However, the differential item functioning detected for this item might be due to the low frequency of the manifestation of internal visualization in this item—nine students expressed internal visualization in the responses given to this item. Another possibility is a potential scoring error for this item, although the interrater reliability coefficient was considerably high for internal visualization ($r = .90$).

Interim Conclusion

Three assumptions about the items regarding internal visualization were tested in the present study. The assumption about item fit was supported by the findings of the study. Item analyses implied that one item might behave differently for each gender. Finally, the assumption about item difficulty did not receive statistical support. The item reliability index (IRIX = .65) indicated that invariant item calibration was not achieved. Therefore, the majority of the item locations on the variable map (Figure 32) are expected to change if another group of second graders take the test.

Findings of the study indicated that each individual test item possessed sufficient quality for measuring internal visualization and that the overwhelming majority of the items could provide appropriate person measures for both genders. As a group, the test items worked efficiently for the majority of the students—person fit values were not estimated for

approximately 90 students in the study because they had a zero score on internal visualization. All the person fit values were between 0.50 and 1.50 for the rest of the students (see Figure 33).

The person reliability index was estimated to be zero in the present study. An indication of this is that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 32) and that the test (with 17 items and without the time limit) did not provide invariant measurement of the internal visualization skill levels of the students. Findings suggested that there might be two discrete item groups with respect to item difficulty, but the variation in the difficulty levels of the items was not high enough to deliver invariant measurement of the trait levels of the students. However, low amount of variation in item difficulty is preferable for the TTCT-figural Form A.

Extending or Breaking Boundaries

The following three assumptions are currently made about the items in Activity 3 regarding extending or breaking boundaries: The items have a good fit, the items are equally difficult, and the items behave the same for students of each gender. In order to test these assumptions, the first six items were analyzed using the dichotomous Rasch model.

Item Fit

The infit mean square values were within the recommended range for a good fit for the first six items in the activity. The infit mean square values suggested that the variation in the item measures was not considerably more or less than the Rasch model projected, meaning that unidimensionality and local independence were not violated. The standardized infit statistics supported this inference. The standardized infit values were also within the recommended range for a good fit for all six items. The standardized infit statistics implied that a good fit was likely

for the first six items in Activity 3. Findings of the study provided statistical evidence for the assumption about item fit.

Item Difficulty

The lower logit point of -0.71 and the upper logit point of 0.69 were estimated for item difficulty—all the items were located between these logit points. The first three items in Activity 3 were all below the 0 logit point in terms of item difficulty, while the other three items were all above the 0 logit point. This is an eye-catching finding. Although each item is essentially the same stimulus, the results implied that the students perceived the items in the first row and the second row differently. The location of a stimulus might be the reason for this difference because the stimuli in the first row have larger space above them than those of in the second row. Another reason could be the distance between the two parallel lines in each pair because the parallel lines in the first row are 0.25 inches apart, while the parallel lines in the second row are 0.75 inches apart.

The strata value of 4.31 was estimated, suggesting that the first six items in Activity 3 measured four different difficulty levels. The item reliability index (IRIX = .90) implied a sufficiently high possibility that these item groups existed and that the locations of the items on the variable map (Figure 34) were determined accurately. Item difficulty analyses suggested that the assumption of the items in Activity 3 being equally difficult regarding extending or breaking boundaries might not be true.

Differential Item Functioning

Analyses yielded different item difficulty levels for each gender. Item difficulty differences ranged from 0.13 to 1.08 logits. Welch's *t*-test results showed that the disparity in item difficulty was statistically significant for the sixth item ($t(85) = -2.13$; $p = .036$), which was

1.08 logits easier for girls. Findings of the study suggested that the sixth item in Activity 3 might not behave the same for students of each gender. However, the differential item functioning detected for this item might be due to the unequal distribution of boys and girls and/or the missing values for this item—the item had the highest number of missing values (53) among all six items included in the analyses. Another reason for this differential item functioning might be a potential scoring error for this item, although the interrater reliability coefficient was .91.

Interim Conclusion

Three assumptions about the first six items in Activity 3 regarding extending or breaking boundaries were examined in the present study. The assumption about item fit received statistical support. The assumption that the items did not exhibit differential item functioning was confirmed for five of the items. Results suggested that the sixth item might behave differently for each gender. Findings of the study did not support the assumption about item difficulty. The sufficiently high item reliability index estimated in the study (IRIX = .90) implied that the locations of the items on the variable map (Figure 34) were determined accurately and that the difficulty levels of the items will remain the same if another group of second graders are administered the test.

Overall, the results indicated that each of the first six items in Activity 3 possessed sufficient quality for measuring extending or breaking boundaries and that the majority of them could provide appropriate person measures for both genders. As a group, the items worked well for the majority of the students—person fit values were not estimated for approximately 50 students because they had either a perfect score or a zero score on extending or breaking boundaries. There were a considerably low number of students with person fit values above 1.50 or below 0.50 (see Figure 35).

The relatively low person reliability index (PRIX = .34) implied that there was some uncertainty in the estimations of the student locations on the variable map (Figure 34) and that the test (with 6 items) did not provide invariant measurement of the extending or breaking boundaries skill levels of the students. There were four groups of items in the activity, but the variation in item difficulty was not high enough to provide invariant measurement of the extending or breaking boundaries skill levels of the students. This was, however, expected considering the number of items included in the analyses. The value for the person reliability index is expected to be much higher when all the items in Activity 3 are included in the analyses.

Humor

Three assumptions are currently made about the items regarding humor. The test items are assumed to have a good fit, to be equally difficult, and to behave the same for boys and girls. In order to investigate these assumptions, the dichotomous Rasch model was utilized.

Item Fit

The infit mean square values were within the suggested range for a good fit for all the items with the exception of the following two items: the item in Activity 1 (I-MNSQ = 1.45) and the ninth item in Activity 2 (I-MNSQ = 1.21). The infit mean square values for 15 good fitting items suggested that the responses given to each of these items were affected by one factor and independent of one another. The standardized infit statistics for these 15 items supported this conclusion. The standardized infit values for the good fitting items were between -2.00 and 2.00, implying that a good fit was highly possible for each of these 15 items.

The infit mean square value for the item in Activity 1 (I-MNSQ = 1.45) implied that there was 45% more variation in the item measures than the Rasch model projected. Forty-five percent more variation indicated that the responses were given in random manner to this item and that a

second factor might have impacted the student responses. The standardized infit value for this particular item (I-STZD = 3.10) suggested that misfit was likely and that a second factor might have in fact impacted the student responses. Because the overwhelming majority of the items on the test were estimated to have a good fit, the potential second factor impacting the responses given to this item was most likely not a construct related to creativity. Therefore, the second factor might be related to the testing conditions specific to this item such as giving a 10-minute time limit for one item and presenting one stimulus on one page. It is also a possible that a scoring error for this item might be the second factor—the interrater reliability coefficient was .85 for humor.

Item fit analyses showed that there was 21% more variation in the item measures than the Rasch model projected for the ninth item in Activity 2 (I-MNSQ = 1.21). Although the amount of variation involved in the item measures for this item denoted some degree of haphazardness in the responses given to this item, the amount of variation was not large enough to denote a second factor according to Adams and Khoo (1996; as cited in Wilson, 2005, p. 129). Additionally, the standardized infit value for this particular item (I-STZD = 0.70) suggested that misfit was not likely.

The study showed that the variation in the item measures was acceptable for 15 of the items. The ninth item in Activity 2 was estimated to involve 21% more variation in the item measures than the Rasch model predicted; however, not only was this variation negligible but also misfit was not likely for this item (as implied by the standardized infit value). On the other hand, analyses showed that the item in Activity 1 involved 45% more variation in the item measures than the Rasch model projected. The standardized infit value for this item implied that the existence of a second factor was likely. Findings of the study supported the assumption about

item fit for 16 items, including the ninth item in Activity 2, and implied that a second factor impacted the responses given to the item in Activity 1.

Item Difficulty

The items were spread between -1.73 and 1.11 logit points with respect to item difficulty. The item in Activity 1 was the easiest item in terms of expressing humor in a response. The items in Activity 2 were either below or above the 0 logit point. Some stimuli in Activity 2 (e.g., the second item; $\delta = -0.92$) generally induced humor in a response, while some other stimuli (e.g., the eighth item; $\delta = 0.64$) did not lead the students to responses with humor expressed. With the exception of the sixth item, the first six items in Activity 3 were all above the 0 logit point in terms of item difficulty. The most difficult item on the test was in Activity 3 (the second item). There was a 1.14-logit difference between the easiest item and the most difficult item in Activity 3. However, only three manifestations of humor caused this difference because humor, in general, was expressed by a few second grade students in the study.

The strata value of 1.57 was estimated, suggesting that all the items represent the same degree of difficulty. The item reliability index was estimated to be .46. This value denoted a moderate chance of the accurate determination of the item locations on the variable map (Figure 36) with regard to item difficulty. Findings suggested that the assumption of the items being equally difficult regarding humor might be true.

Differential Item Functioning

The Rasch model yielded unequal item difficulty levels for boys and girls. Differences in item difficulty ranged from 0.13 to 1.78 logits. None of the differences in item difficulty was statistically significant at the $p < .05$ level according to Welch's *t*-test results. Findings showed

that all the test items behaved the same for boys and girls with respect to humor. The study provided supporting evidence for the assumption about differential item functioning.

Interim Conclusion

Three assumptions about the items regarding humor were tested in the present study. The assumption that the items have a good fit was confirmed for 16 of the items. Findings implied that the responses given to the item in Activity 1 were impacted by a second factor, but this second factor was most likely related to the testing conditions or a potential scoring error specific to this item. The assumptions about differential item functioning and item difficulty also received statistical support. A moderately high item reliability index was estimated in the study (IRIX = .46). Because the item reliability index was below the cutoff point recommended for invariant item calibration, the majority of the item difficulties (i.e., item locations on the variable map [Figure 36]) are expected to change if another group of second graders are administered the test.

Findings of the study indicated that the overwhelming majority of the items possessed good quality for measuring humor and providing appropriate person measures for both genders. As a group, the items worked efficiently for approximately 60 of the students—person fit values were not estimated for over 100 students because they did not express humor in any item. For the rest of the students, all of them had person fit values between 0.50 and 1.50 (see Figure 37).

A zero value was estimated for the person reliability index because there was essentially one item group with respect to item difficulty. This indicated that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 36) and that the test (with 17 items and without the time limit) did not provide invariant measurement of the

humor skill levels of the students. No variation in item difficulty, however, is desired for the TTCT-figural Form A.

Richness of Imagery

Rasch measurement theory analyses were conducted to test the following three assumptions about the items regarding richness of imagery: The items have a good fit, the items are at the same difficulty level, and the items behave the same for each gender.

Item Fit

The infit mean square values were within the suggested range for a good fit for all the test items. The infit mean square statistics showed that the variation in the item measures was acceptable for all the items. This result suggested that the responses given to each item were neither too haphazard nor too deterministic. The standardized infit statistics confirmed this conclusion. The standardized infit values were between -2.00 and 2.00, implying that a good fit was highly possible for all the test items. Findings provided supporting evidence for the assumption about item fit.

Item Difficulty

Item difficulties ranged from -2.38 to 0.97 logits. The majority of the items were clustered between -0.73 and 0.97 logit points. The item in Activity 1 was the easiest in terms of creating a memorable figural response. The items in Activity 2 were either below or above the 0 logit point. It was easier for the students to create a memorable figural response using some stimuli in Activity 2 (e.g., the fourth item; $\delta = -0.73$). Some other stimuli (e.g., the tenth item; $\delta = 0.57$) was relatively challenging in terms of expressing richness of imagery in a response. The first six items in Activity 3 were all above the 0 logit point with respect to difficulty. It was usually difficult for the students to create a memorable figural response in Activity 3. There was

a 0.91-logit difference between the easiest item and the most difficult item in Activity 3—14 manifestations of richness of imagery led to the estimation of this difference.

The strata value of 4.06 was estimated. This value suggested that there were four distinct item groups on the test with respect to item difficulty. The item reliability index estimated in the study (IRIX = .89) denoted a considerably high chance of the presence of these item groups and the accurate determination of the locations of the items on the variable map (Figure 38). The assumption of the items being equally difficult regarding richness of imagery was not supported by the findings.

Differential Item Functioning

Gender-based item analyses yielded different item difficulty levels for boys and girls. Differences in item difficulty were between 0.08 and 1.24 logits. Welch's *t*-test results revealed that none of the differences in item difficulty was statistically significant at the $p < .05$ level. Results indicated that all the test items behaved the same for boys and girls regarding richness of imagery. The assumption that the items do not favor students of one gender over students of the opposite gender received statistical support.

Interim Conclusion

Three assumptions about the items regarding richness of imagery were investigated in the present study. The assumptions about item fit and differential item functioning were confirmed. The assumption about item difficulty, however, did not receive statistical support. The considerably high item reliability index (IRIX = .89) suggested that invariant calibration of the items was most likely achieved. The difficulty levels of the items (i.e., item locations on the variable map [Figure 38]) are expected to remain the same if another group of second graders take the test.

Findings of the study indicated that each test item possessed sufficient quality for measuring richness of imagery and providing appropriate person measures for both genders. As a group, the items worked as predicted by the Rasch model for the majority of the students—person fit values were not estimated for approximately 35 students because they had a zero score on richness of imagery. For the rest of the students, nearly all the person fit values were between 0.50 and 1.50 (see Figure 39).

Although there were four different item groups regarding item difficulty, the person reliability index was estimated to be .20. An indication of this finding is that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 38) and that the test (with 17 items and without the time limit) did not provide invariant measurement of the richness of imagery skill levels of the students. Because the TTCT-figural Form A is timed test, low amount of variation in item difficulty is not an issue—it is in fact preferable.

Colorfulness of Imagery

Three assumptions are currently made about the items regarding colorfulness of imagery. The items are assumed to have a good fit, to be equally difficult, and to behave the same for students of each gender. In order to test these assumptions, the dichotomous Rasch model was utilized.

Item Fit

The infit mean square values were between 0.80 and 1.20 for all the items, meaning that the variation in the item measures was acceptable with regard to colorfulness of imagery. The infit mean square statistics implied that unidimensionality and local independence were not violated. These inferences were supported by the standardized infit statistics. The standardized

infit values were within the recommended range for a good for all the items, indicating that a good fit was highly possible for each item. Findings of the study provided supporting evidence for the assumption that the items have a good fit regarding colorfulness of imagery.

Item Difficulty

The difficulty levels of the items ranged from -2.72 to 1.20 logits, with the majority of the items being between -0.30 and 0.49 logits. The item in Activity 1 was the easiest in terms of creating an exciting figural response. The items in Activity 2 were either below or above the 0 logit point. It was relatively easy for the students to create an exciting figural response using some stimuli in Activity 2 (e.g., the third item; $\delta = -1.52$). On the other hand, some other stimuli (e.g., the tenth item; $\delta = 1.14$) did not lead the students to create exciting figural responses as often as the other stimuli did. The first six items in Activity 3 were all above the 0 logit point with the exception of the third item. It is usually difficult for the students to create an exciting figural response in Activity 3. There was a 1.50-logit difference between the easiest item and the most difficult item in Activity 3, suggesting that the items in Activity 3 might not be at the same difficulty level for colorfulness of imagery. A possible scoring error, however, may have caused this disparity—the interrater reliability coefficient was .86 for colorfulness of imagery.

The strata value was estimated to be 4.49. This value suggested that there were four distinct item groups on the test with respected to item difficulty. The sufficiently high item reliability index estimated in the study (IRIX = .91) implied that these item groups existed and that the locations of the items on the variable map (Figure 40) were determined accurately. Item difficulty analyses did not support the assumption that the items are equally difficult regarding colorfulness of imagery.

Differential Item Functioning

Item analyses yielded different item difficulties for each gender. The range for the differences in item difficulty was from 0 to 0.95 logits. According to Welch's *t*-test results, none of the differences in item difficulty was statistically significant at the $p < .05$ level. Findings of the study showed that the items did not favor students of one gender over students of the opposite gender. The present study provided supporting evidence for assumption that the items behave the same for students of each gender.

Interim Conclusion

Three assumptions about the items regarding colorfulness of imagery were tested in the present study. The assumptions about item fit and differential item functioning were supported by the findings of the study. The study did not provide supporting evidence for the assumption that the items are the same difficulty level. The item reliability index (IRIX = .91) denoted a sufficiently high chance of the presence of the four items groups and the accurate determination of the item locations on the variable map (Figure 40). Therefore, item difficulties will remain at the same level if another group of second graders are administered the test.

Findings of the study indicated that each individual test item possessed sufficient quality for measuring colorfulness of imagery and providing appropriate person measures for both genders. As a group, the items worked as predicted by the Rasch model for the majority of the students. There were a considerably low number of students with person fit values above 1.50 (see Figure 41). There were several students with person fit values below 0.50; however, they indicated that the Rasch model predicted perfectly the kind of responses those students would give. It should be noted that person fit values were not estimated for approximately 30 students because they had a zero score on colorfulness of imagery.

A value of .11 was estimated for the person reliability index. This value indicated that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 40) and that the test (with 17 items and without the time limit) did not provide invariant measurement of the colorfulness of imagery skill levels of the students. If the students were given the TTCT-figural Form B under the same testing conditions as the present study, the overwhelming majority of the student locations on the variable map would change. Although there were four discrete item groups, the variation in item difficulty was not as much as the Rasch model projected. However, low amount of variation in item difficulty is desired for the TTCT-figural Form A.

Fantasy

The following three assumptions are currently made about the test items regarding fantasy: The items have a good fit, the items are at the same difficulty level, and the items do not favor students of one gender over students of the opposite gender. These assumptions were examined utilizing the dichotomous Rasch model.

Item Fit

The infit mean square values were within the suggested range for a good fit for 13 of the items. The following four items showed misfit: the sixth item (I-MNSQ = 0.79) as well as the ninth item (I-MNSQ = 1.22) in Activity 2, and the second item (I-MNSQ 0.78) as well as the sixth item (I-MNSQ 0.79) in Activity 3. The infit mean square values for 13 good fitting items showed that there was acceptable variation in the item measures for each of these 13 items. Findings implied that a single factor affected the responses given to the items and that the responses given to these items were not too deterministic. These inferences were supported by the standardized infit statistics. The standardized infit values were between -2.00 and 2.00 for

these 13 items. The standardized infit statistics suggested that a good fit was highly possible for each good fitting item.

The infit mean square value for the ninth item in Activity 2 (I-MNSQ = 1.22) showed that there was 22% more variation in the item measures than the Rasch model predicted for this item. This variation suggested that the students randomly manifested fantasy in this item. In other words, some students who manifested fantasy in this item were not predicted to express fantasy in their responses given to this item (and *vice-versa*). This random expressions of fantasy implied that a second factor might have impacted the student responses. The standardized infit value for this item (I-STZD = 1.10), however, suggested that the existence of a second factor was not likely. In addition to this, this 22% more variation can be neglected because Adams and Khoo (1996; as cited in Wilson, 2005, p. 129) argued that up to 33% more variation than Rasch measurement theory predicted did not degrade person measures.

The infit mean square values for the sixth items in Activity 2 (I-MNSQ = 0.79) and Activity 3 (I-MNSQ = 0.79) denoted 21% less variation in the item measures than the Rasch model predicted. Similarly, the infit mean square value for the second item in Activity 3 (I-MNSQ 0.78) indicated that the variation in the item measures was 22% lower than the variation projected by the Rasch model. The infit mean square values for these three items suggested that the responses given to each of these items were too deterministic. The standardized infit values for these three items, however, showed that overfit was not likely for any of these items—the standardized infit values were -0.70, -0.60, and -0.60 respectively. Additionally, the amount of variation in these three items was negligible because Adams and Khoo (1996; as cited in Wilson, 2005, p. 129) argued that up to 25% less variation in the item measures than the Rasch model predicted did not degrade person measures.

The study showed that the variation in the item measures was acceptable for 13 of the items. The infit mean square values for the sixth item as well as the ninth item in Activity 2, and the second item as well as the sixth item in Activity 3 indicated that the amount of variation in the item measures was beyond the amount of acceptable variation recommend in the literature (Bond & Fox, 2001; Engelhard, 2013). However, the standardized infit values for each of these four misfitting items implied that misfit was not likely for any of these items. Findings of the study confirmed the assumption about item fit for 13 items and implied that the four misfitting items had an acceptable fit.

Item Difficulty

The lower logit point of -3.13 and the upper logit point of 0.99 were estimated for item difficulty. The majority of the items were located between -0.93 and 0.99 logit points. The item in Activity 1 was the easiest in terms of expressing fantasy in a response. The items in Activity 2 were either below or above the 0 logit point. Some stimuli in Activity 2 (e.g., the fourth item; $\delta = -0.93$) induced fantasy in a response more often than some other stimuli did (e.g., the tenth item; $\delta = 0.58$). With the exception of the fifth item, the first six items in Activity 3 were all above the 0 logit point in terms of item difficulty, implying that expressing fantasy in a response in Activity 3 was relatively difficult. The most difficult item on the test was in Activity 3 (the first item). There was a 0.92-logit difference between the easiest item and the most difficult item in the activity—only five manifestations of fantasy led to the estimation of this difference.

The strata value of 3.23 was estimated. This value denoted the presence of three different item groups on the test with respect to item difficulty. The considerably high item reliability index estimated in the study (IRIX = .83) suggested that these item groups possibly existed and that the difficulty levels of the items (i.e., item locations on the variable map [Figure 42]) were

most likely determined accurately. Item difficulty analyses did not support the assumption that the items are equally difficult regarding fantasy.

Differential Item Functioning

Gender-based item analyses estimated unequal item difficulty levels for boys and girls. Differences in item difficulty ranged from 0.04 to 1.93 logits. Welch's *t*-test analyses showed that none of the differences in item difficulty was statistically significant at the $p < .05$ level. Findings showed that the items in the TTCT-figural Form A behaved the same for each gender regarding fantasy. The study provided supporting evidence for the assumption about differential item functioning.

Interim Conclusion

Three assumptions about the items regarding fantasy were investigated in the present study. The assumptions about item fit and differential item functioning received supporting evidence from the study. The assumption about item difficulty was not supported. Because the item reliability index was .07 below the cutoff point recommended for invariant item calibration, there is a possibility that the locations of a few of the items on the variable map (Figure 42) will change if another group of second graders are administered the test.

Findings of the study indicated that the test items possessed sufficient quality for measuring fantasy and providing appropriate person measures for both genders. As a group, the items did not work as well as the Rasch model projected. There were several students with person fit values above 1.50 (see Figure 43), indicating that these students did not manifest fantasy in a response when the Rasch model predicted them to do so (and *vice-versa*). It should be noted, however, that one or two expressions of fantasy led to the estimation of person fit

values above 1.50. Another important point is that person fit values were not estimated for approximately 90 students because they had a zero score on fantasy.

The person reliability index was estimated to be zero. This finding indicated that there was considerable uncertainty in the estimations of the student locations on the variable map (Figure 42) and that the test (with 17 items and without the time limit) did not provide invariant measurement of the fantasy skill levels of the students. Therefore, if the students were given the TTCT-figural Form B under the same testing conditions as the present study, the locations of the students on the variable map would change. Although there were four different item groups with respect to item difficulty, the variation was not as much as the Rasch model predicted. However, zero variation in item difficulty is desired for the TTCT-figural Form A.

CHAPTER 6

CONCLUSION

Voluminous research has been conducted on both forms of the TTCT-figural since the tests were first published in 1966. Although there is considerable research on the reliability, validity, and fairness of the total test scores (e.g., Matud & Grande, 2007; Torrance, 1981, 1998), no study has explored the quality of the test items. Currently, there is no statistical data on item fit, item difficulty, average item difficulty, item scaling, differential item functioning, and item information.

Due to the absence of research, certain assumptions are made about the items in the TTCT-figural. The items are assumed to have a good fit, to be equally difficult, to behave the same for examinees of each gender, and to involve the same amount of standard error of measurement (i.e., to provide the same amount of information on a particular norm-referenced variable). Additionally, it is assumed that the response options of the items (or activities) are distributed evenly on the latent trait continuum regarding the polytomously scored variables.

None of these assumptions has been yet examined using appropriate methods. Because there is no supporting evidence, these assumptions raise methodological concerns. The purpose of the present study was to conduct Rasch measurement theory analyses on the test items for each of the 18 variables measured in the TTCT-figural Form A as well as for elaboration-I to examine the assumptions about item fit, item difficulty, average item difficulty, item scaling, differential item functioning, and item information.

In order to test these assumptions, four main research questions were identified. The first research question investigated item fit; the second research question addressed item difficulty, average item difficulty, item scaling, and differential item functioning; the third research question explored the standard error of measurement involved in each item; and the final research question examined invariant item calibration as well as invariant measurement of the trait levels of the students.

First Research Question

The first research question investigated item fit for the test items regarding each individual variable measured in the TTCT-figural Form A as well as elaboration-I. Specifically, the first research question addressed whether a particular variable was the only factor affecting the responses given to the items for that variable and whether the responses given to the items were independent of one another.

Infit mean square and standardized infit statistics were utilized to determine whether the items had a good fit or misfit with regard to a particular variable. Regarding nine variables, both the infit mean square and standardized infit values were within the recommended ranges for all the items included in the analyses (i.e., all the items had a good fit). Regarding five variables, the infit mean square values for some items were either below 0.80 or above 1.20, but the standardized infit values for those items indicated that misfit was not likely. Regarding three variables, both the infit mean square and standardized infit values were out of the suggested ranges for a good fit for one item. In other words, the infit mean square value for the item implied that the item had misfit, and the standardized infit value indicated that misfit for the item was in fact possible.

Both the infit mean square and standardized infit values were within the recommended ranges for a good fit for all the items regarding the following variables: originality, elaboration, resistance to premature closure, movement or action, unusual visualization, internal visualization, extending or breaking boundaries, richness of imagery, and colorfulness of imagery. Regarding these variables, none of the items on the test involved significantly more or less variation than Rasch measurement theory projected. In other words, a single factor affected the responses given to the items (i.e., responses were not given haphazardly), and the responses given to the items were independent of one another (i.e., responses were not too deterministic).

Regarding the following variables, the infit mean square value for at least one item was below 0.80 or above 1.20, but the standardized infit values were between -2.00 and 2.00 for all the items: fluency, abstractness of titles, emotional expressiveness, storytelling articulateness, and fantasy. An infit mean square value above 1.20 denoted the existence of a second factor impacting the responses given to the item, while an infit mean square value below 0.80 implied that the responses given to the item were too deterministic. In either case, the standardized infit value indicated that misfit (i.e., a second factor or the predictability of the responses) was not likely for the item and that the variation in the item measures was acceptable (i.e., not detrimental). For instance, regarding fluency, the infit mean square value for the item in Activity 1 (I-MNSQ = 1.29) showed underfit, but the standardized infit value (I-STZD = 1.10) suggested that misfit was not likely.

Both the infit mean square and standardized infit values were out of the recommended ranges for one item regarding the following variables: elaboration-I, expressiveness of titles, and humor. Regarding each of these variables, one item on the test had significantly more or less variation than Rasch measurement theory projected for this item. In other words, the amount of

variation in the item measures threatened unidimensionality or local independence for this item. For instance, regarding expressiveness of titles, the infit mean square value for the item in Activity 1 (I-MNSQ = 1.35) denoted underfit, and the standardized infit value (I-STZD = 3.60) indicated that overfit for this item was in fact possible. Underfit for this item implied that a second factor impacted the responses given to this item. However, this second factor was related to either the testing conditions specific to this particular item (e.g., giving a 10-minute time limit for one item and presenting one item on one page) or a scoring error.

On the other hand, regarding elaboration-I, the infit mean square value for the second item in Activity 3 denoted overfit (I-MNSQ = 0.75), and the standardized infit value for the same item (I-STZD = -2.40) indicated that overfit for this item was in fact likely. Overfit for the second item in Activity 3 suggested that one could predict how many details a second grade student would approximately add to his or her response to this particular item.

Overall, analyses of item fit showed that there was acceptable variation in the item measures for the overwhelming majority of the items. Findings of the study suggested that each individual test item possessed sufficient quality for providing appropriate person measures regarding the overwhelming majority of the variables. As can be seen in Figure 45, a few of the infit mean square values were below 0.80 or above 1.20. Additionally, only three of the items had standardized infit values below -2.00 or above 2.00 (see Figure 46).

Second Research Question

The second research question explored the difficulty levels of items (or activities for elaboration) regarding each individual variable measured in the TTCT-figural Form A as well as elaboration-I. The second question also addressed the assumption about differential item functioning.

Regarding the dichotomously scored variables, the question addressed whether each test item was equally difficult to manifest a particular dichotomously scored variable. The difficulty levels of the items were estimated for each variable. For the majority of the variables, the first item was estimated to be the easiest item on the test. This was most likely because the students had plenty of time to draw different characters or objects and to manifest several variables in their responses given to this particular item. For some variables (e.g., storytelling articulateness and unusual visualization), the tenth item in Activity 2 was the most difficult item. For some other variables (e.g., fluency and originality), one of the items in Activity 3 was estimated to be the most difficult.

The first six items in Activity 3 were estimated to have different difficulty levels regarding all the variables measured by the items, although each item (i.e., parallel line) in the activity is essentially the same stimulus. Regarding the majority of the variables, the difference was large enough to imply that the items were not at the same difficulty level. Findings suggested that the items in Activity 3 might not be equally challenging (i.e., difficult) in terms of expressing a particular creative thinking skill.

In addition to item difficulties, Facets (Linacre, 2011) also provided strata values. Strata values were utilized to investigate whether the items differed from each other with respect to difficulty or whether the items were at the same difficulty level. Strata values indicated that the items were not at the same difficulty level regarding the overwhelming majority of the variables. For instance, regarding originality, there were six item groups. Regarding richness of imagery, on the other hand, there were four item groups. Only the strata value for humor denoted one item group in terms of item difficulty.

Regarding the polytomously scored variables, the second research question investigated whether all the test items were at the same average difficulty level and whether the response options of the items were based on equal intervals (regarding elaboration-I, whether every additional detail added to a response required the same amount of increase in the trait level). In order to answer the second research question, average difficulty levels of the items (or activities for elaboration), strata values, the Rasch-Andrich thresholds, and the Rasch-Thurstone thresholds were estimated.

Analyses yielded different average difficulty levels for the items (or activities). In order to test whether the items (or activities) were at the same average difficulty level, the strata values were considered. The strata value for elaboration-I implied that all the items were within the same group with respect to difficulty. The strata values for the other variables indicated that there were at least two item (or activity) groups in terms of average item (or activity) difficulty. For instance, the results indicated that there were three item groups regarding resistance to premature closure—these three item groups (with the 3-point scale currently used to score resistance to premature closure) measured five distinct difficulty levels.

The Rasch-Andrich thresholds were estimated for the items (or activities for elaboration) to explore whether the response options were based on equal intervals. However, because disordered Rasch-Andrich thresholds were estimated for the items regarding all the polytomously scored variables, the Rasch-Thrustone thresholds were utilized to test whether the response options were equally distributed on the latent trait continuum. The Rasch-Thrustone thresholds were used because Linacre (2010) argued that the Rasch-Thrustone thresholds should be used to explore whether the polytomously scored items on a test were based on equal intervals when the Rasch-Andrich thresholds exhibited disordering.

The Rasch-Thrustone thresholds estimated for the items (or activities for elaboration) indicated that the response options of the items (or activities) were not distributed evenly on the latent trait continuum regarding elaboration, elaboration-I, and abstractness of titles. Findings suggested that the response options of the items were distributed evenly regarding resistance to premature closure.

As can be seen in Figure 47, the difficulty level (regarding the dichotomously scored variables) or average difficulty level (regarding the polytomously scored variables) of the item in Activity 1 was always below the 0 logit point. This finding indicated that it was generally easy for the students to manifest a particular variable in the responses given to this item. With the exception of the tenth item, all the items in Activity 2 were either below or above the 0 logit point. It was relatively challenging for the students to express a particular variable in the responses given to the tenth item in Activity 2. Finally, all the items in Activity 3 were either below or above the 0 logit point.

In order to investigate whether the items on the test behaved the same for each gender, the difficulty levels of the items were estimated for each gender and compared using Welch's *t*-test. Welch's *t*-test results showed that the difficulty levels of the items were not significantly different for boys and girls for any of the items regarding nine variables, meaning that the test items behaved the same for each gender regarding those variables. Regarding three of the variables, differential item functioning was detected for only one item. Regarding four of the variables, differential item functioning was detected for two items. Regarding one variable, three items were estimated to behave in a significantly different manner for boys and girls.

One item, the seventh item in Activity 2, was estimated to behave differently regarding three variables. Differential item functioning was detected for the seventh item in Activity 2

regarding resistance to premature closure, storytelling articulateness, and expressiveness of titles. The item was easier for boys regarding all three variables.

It should be noted that the differential item functioning detected for the items might not be due to the actual performance differences regarding some variables. Regarding some variables, the detected differential item functioning may have been because of the low frequency of the manifestation of the variable, unequal distribution of boys and girls in the study, the missing values for these items, and/or a potential scoring error.

Third Research Question

The third research question examined the standard error of measurement involved in item scores for each norm-referenced variable. The question addressed whether the standard error of measurement involved in each individual item (or activity) was the same across the items (or activities) regarding each norm-referenced variable. The question also addressed whether the standard error of measurement involved in each individual item (or each of the three activities) was at the same level along the latent trait continuum.

In order to answer this research question, item information functions were estimated for the norm-referenced variables. Analyses of item information demonstrated that each individual test item (or activity) provided that same amount of information on a variable. However, because the items (or activities) differed from each other in terms of item difficulty or average item difficulty, each item (or activity) delivered this information within a different logit range on the latent trait continuum.

Two types of item information functions were obtained in the study. The first type of item information functions had bell shapes. Bell-shaped item information functions were estimated for the dichotomously scored variables as well as for resistance to premature closure.

The bell-shaped item information functions for the dichotomously scored variables were anticipated because according to Rasch measurement theory, dichotomously scored items have bell-shaped information functions.

The bell-shaped information functions estimated for the items regarding resistance to premature closure, however, were unexpected. This is because the items were scored in a polytomous manner regarding resistance to premature closure and thus were projected to have item information functions with multiple peaks. The rating scale model analyses showed why the items had bell-shaped item information functions regarding resistance to premature closure. As previously mentioned, the intermediate response option of the items measured a narrow section of the variable regarding resistance to premature closure. In other words, the distance between the first threshold and the second threshold was not large enough to yield a wide item information function. Additionally, the intermediate response option provided the highest amount of information at an item's average difficulty point. Therefore, bell-shaped item information functions were obtained for the test items.

The bell-shaped item information functions implied that each item provided the highest amount of information (i.e., involved the lowest amount of standard error of measurement) at the item's difficulty point (or average difficulty point for resistance to premature closure) on the latent trait continuum. The bell-shaped item information functions also indicated that the amount of item information declined as the information function approached the far ends of the latent trait continuum. This finding implied that the amount of standard error of measurement increased towards the far ends of the latent trait continuum.

The second type of item information functions were wide and had more than one peak. The second type of information functions were obtained for elaboration and abstractness of titles.

Wide item information functions with multiple peaks were observed for the items (or activities) regarding these variables because at least one of the intermediate response options of the items measured a wide section of the variable. For instance, the second response option of the items measured a narrow section of the variable regarding abstractness of titles, but the third response option measured a wide section of the variable. Thus, a wide item information function with two peaks was estimated for each test item regarding abstractness of titles.

The amount of information provided by the items on elaboration and abstractness of titles declined in the direction of the far ends of the latent trait continuum. This finding suggested that the amount of standard error of measurement increased towards the far ends of the latent trait continuum for the items regarding elaboration and abstractness of titles.

Analyses of item information showed that each test item in the study offered the same amount of information on a particular variable and that the amount of information provided by an item was not at the same level along the latent trait continuum. Whether it was estimated for a dichotomously scored variable or a polytomously scored variable, the information function of an item declined towards the far ends of the latent trait continuum. This incident implied that the amount of standard error of measurement involved in the total scores on a norm-referenced variable was not the same for all scores on this particular variable (i.e., all scores on a norm-referenced variable were not equally reliable).

The test information function estimated for each norm-referenced variable supported the conclusion that not all scores on a norm-referenced variable were equally reliable. As seen in Figure 44, the test information function for each variable decreased as the information function approached the far ends of the latent trait continuum. Findings suggested that when the test is

administered under the standard testing conditions, some scores on a particular variable will be more reliable.

Fourth Research Question

The final research question investigated invariant item calibration and invariant measurement of the trait levels of second grade students. The question tested whether the items (or activities for elaboration) in the TTCT-figural Form A could be calibrated in a sample-independent manner so that the difficulty levels of the items (or activities) will remain the same whenever another group of second graders take the test. The fourth research question also tested whether the creative thinking skill levels of the students could be measured in an item-independent manner by the TTCT-figural Form A so that the trait levels of second grade students would remain at the same level if they were given another set of items measuring the same construct (e.g., TTCT-figural From B).

In order to answer the first part of the question, which was about invariant item calibration, an item reliability index was estimated for each variable. An item reliability index value of or above .90 was estimated for seven of the variables—note that .90 was the cutoff point suggested for invariant item calibration (Linacre, 2016). Regarding these seven variables, the results indicated that the locations of the items on variable map were determined accurately and that the items will keep their relative locations if another group of second graders are administered the test. For the rest of the variables, an item reliability index value below .90 was estimated. The item reliability indexes were between .80 and .90 for seven variables. Although those item reliability index values (the ones between .80 and .90) were below the cutoff value suggested for invariant item calibration, they denoted a considerably high possibility of the accurate determination of the difficulty levels of the items. The item reliability indexes were

below .70 for three variables. Regarding these three variables, the item reliability indexes implied that the difficulty levels of the majority of the items most likely will change if another group of second graders take the test.

In order to answer the second part of the question, which concerned invariant measurement of the trait levels of the students, person reliability indexes were estimated. None of the person reliability indexes were above the cutoff point suggested for invariant measurement of the trait levels of the students, .80 (Linacre, 2016). The highest person reliability index was estimated for elaboration-I (PRIX = .75). The person reliability index value of 0 was estimated for several dichotomously scored variables.

Considerably low values or zero values for the person reliability indexes implied that the variation in the difficulty levels of the items or the difficulty levels of the response options of the items was not high enough to yield invariant measurement of the trait levels of the students. A small amount of variation—in fact no variation—is preferable for the items in the TTCT-figural Form A with respect to item difficulty or average item difficulty because it is a timed test. However, the person reliability index should have been much higher for the polytomously scored variables. This is because the polytomously scored variables were developed to capture various trait levels in one item.

Implications

The findings of the present study had certain implications. These implications concern both practitioners who administer the TTCT-figural Form A for educational purposes and the Scholastic Testing Service, the company responsible for standardizing the test and publishing the scoring manual.

Item fit analyses showed that the items generally worked as projected by Rasch measurement theory. Analyses of item fit indicated that the responses given to the items were neither too haphazard nor too deterministic regarding the overwhelming majority of the variables. In other words, the variation in the item measures was acceptable. This finding provided evidence for construct validity of the TTCT-figural Form A. Results implied that the variation in the item in Activity 1 may degrade the measurement of expressiveness of titles skills and humor skills to a certain extent.

Analyses of item difficulty showed that the items were not at the same difficulty level regarding the overwhelming majority of the variables—the items were at the same level regarding only humor. Findings of item difficulty suggested that each item should not receive the same score for the manifestation of a particular variable. This finding especially concerns fluency and originality because each test item receives a score when the items are scored for each of these two variables. However, this finding also concerns the creative strengths. Further research should be conducted to examine the difficulty levels of the items regarding the dichotomously scored variables, and experts should determine the appropriate scores for each item based on its difficulty level.

Findings of average item (or activity for elaboration) difficulty showed that the items (or activities) were not at the same average difficulty level regarding elaboration, abstractness of titles, and resistance to premature closure—the items were within the same group regarding elaboration-I. Additionally, the results indicated that the response options of the items were not based on equal intervals regarding the all the polytomously scored variables with the exception of resistance to premature closure.

The response options of the items not being equally distributed is not an important issue regarding item level elaboration (i.e., elaboration-I) because the items were within the same group with respect to average item difficulty. Thus, the same number of details added to a response corresponds to the same trait level across the items. However, the response options of the items not being equally distributed causes a problem regarding elaboration, abstractness of titles, and resistance to premature closure. This is because the results showed that the same type of response did not correspond to the same trait level regarding these three variables, as assumed.

Further research should be conducted to examine the average difficulty levels of the items (and activities for elaboration) as well as the difficulty levels of the response options of the items regarding the polytomously scored variables. Based on this research, experts should define the trait levels that correspond to the response options of the items regarding a particular variable and assign appropriate scores to each response option based on the average difficulty level of the item as well as the difficulty level of the response option.

Analyses of item information implied that not all scores on a particular variable are equally reliable and that not all the trait levels were measured with equal precision. After further research is conducted following the standard testing procedures, a group of experts should decide how reliable each score on a particular variable is.

Finally, analyses of invariant measurement of the trait levels of the students implied that the rating scales which were used to score the items regarding the polytomously scored variables did not work efficiently for students at this grade level—the person reliability index values were below .80 for all the polytomously scored variables. Further research should be conducted to investigate the reasons for this situation. This issue may be resolved by revising the rating scale for some variables. For instance, a group of experts can revise the scaling for elaboration and

determine in what way the scaling for elaboration would work better for students at this grade level. This may be achieved by changing the number of details that correspond to a particular response option.

Limitations of the Study

The TTCT-figural Form A is a different kind of test from the ones that are typically analyzed with Rasch measurement theory (i.e., instruments that require examinees to give the right answer to a question or to choose one of the predetermined response options). Due to the unique nature of the TTCT-figural Form A, a few adjustments had to be made in the present study. For the present study, adjustments were made in terms of test administration (e.g., giving a supplementary time for Activity 2) and data analyses (e.g., creating a 6-point scale for elaboration-I). These were necessary adjustments for the present study, but it should be noted that they were not parts of the standard testing or scoring procedures.

The sample size of the study was large enough to provide interpretable results for the overwhelming majority of the variables. However, because synthesis of incomplete figures and synthesis of lines were manifested by a considerably small number of students in the study, no meaningful results were obtained for the items regarding these two variables. Findings indicated that synthesis of incomplete figures and synthesis of lines should be investigated with a much larger sample size.

Meaningful and interpretable results were obtained for the items regarding the rest of the variables measured in the TTCT-figural as well as elaboration-I. However, because the overwhelming majority of the variables were not expressed by many students, a few differences in the number of manifestations of a variable led to the estimations of considerably different item

difficulty levels. Therefore, to eliminate this issue, the items in TTCT-figural Form A should be examined with a larger sample size.

The item in Activity 1 and all 10 items in Activity 2 were analyzed in the present study, but only the first six items in Activity 3 were examined regarding the variables measured by the items. Findings of the study suggested that the location of a parallel line pair and/or the distance between the parallel lines in each pair might impact the item parameters regarding some variables. Therefore, all 30 items in Activity 3 should be included in the analyses to examine whether each test item in Activity 3 is perceived the same way by the examinees and whether each item in the activity was at the same difficulty level.

The items were scored based on the scoring manual published in the U.S. (Ball & Torrance, 1984). This situation is not an issue for the overwhelming majority of the variables, but it impacts the scoring for originality. Whether a response was original was determined using the list which was created utilizing the responses given by students in the U.S. It is possible that a different list would be obtained if the TTCT-figural Form A were standardized in Turkey. The test items should be analyzed using the Turkish norms for originality in the future.

The majority of the interrater reliability coefficients indicated that there were not many scoring errors that might have impacted the item parameters estimated in the present study. The majority of the interrater reliability coefficients were at or above .89. However, the interrater reliability coefficients for elaboration-I (r = .84), unusual visualization (r = .87), humor (r = .85), richness of imagery (r = .87), and colorfulness of imagery (r = .86) were relatively lower. It is possible that a potential scoring error may have affected the estimations of item parameters regarding some variables. Thus, the TTCT-figural Form A should be analyzed using the tests that were scored by the experts at the Scholastic Testing Service.

When the tests are scored by the experts at the Scholastic Testing Service, the interrater reliability coefficients will be possibly over .90 for all the variables. However, having interrater reliability coefficients over .90 indicates that the severity of a rater still systematically impacts the item parameters and the examinee scores. In order to examine the impact of the severity of a rater on item parameters, the items should be analyzed using the many-facets Rasch model. Many-facets Rasch model takes the severity of a rater into account and can reveal how much discrepancy exists in item difficulty levels across the raters.

Directions for Future Research

This study is the first of its kind. No study has examined the items in the TTCT-figural Form A (and Form B) using Rasch measurement theory. Future research that utilizes Rasch measurement theory to examine the items in TTCT-figural Form A (and Form B) is needed.

Future research should analyze the test items for all grade levels to test whether the seven assumptions addressed in the present study are supported by sound evidence at all grade levels. In addition to analyzing differential item functioning for gender, the difficulty levels of the items should be estimated for students from different ethnicities, socioeconomic status, and creativity fields (e.g., art, music, science, and literature). Therefore, differential item functioning can be investigated from different aspects. Finally, the test items should be examined in more detail regarding all the variables measured in the test.

The present study critiqued the test items regarding all the variables measured in the TTCT-figural as well as elaboration-I for several parameters (e.g., item fit, item difficulty, average item difficulty, item scaling, differential item functioning, item information, invariant item calibration, and invariant measurement of the trait levels of the students). However, the findings on all these parameters should be scrutinized for each individual test item in more detail.

Therefore, the reasons for misfit, for the variation in item difficulty as well as average item difficulty, for unequal distribution of the response options, and for differential item functioning can be better understood. Thereby, effective solutions can be offered.

Concluding Remarks of the Dissertation

The present study addressed seven assumptions about the items in the TTCT-figural Form A by answering four main research questions. The study investigated the items in the TTCT-figural Form A for the following item parameters: item fit, item difficulty, average item difficulty, item scaling, differential item functioning, item information, invariant item calibration, and invariant measurement of the trait levels of second grade students. Because the CTT and factor analysis frameworks cannot address all seven assumptions and answer all the research questions of the present study, the Rasch measurement theory framework was utilized.

The dichotomous Rasch model was used to analyze the test items regarding the dichotomously scored variables. The rating scale model was used to investigate the items (or activities) regarding the polytomously scored variables. Rasch measurement theory analyses confirmed some of the assumptions about the items, but not all assumptions received statistical support. Overall, the present study showed that the TTCT-figural Form A worked considerably well regarding the majority of the variables measured in the test, although the test was developed using the CTT framework.

Findings of the present study suggested that further research should be conducted to examine the assumptions about the test items. A group of experts, then, should analyze the findings of future research and make decisions about the test items. This way, the TTCT-figural will measure the trait levels of students more precisely and provide more valid test scores.

REFERENCES

Acar, S. (2013). *Empirical studies of literal divergent thinking* (Doctoral dissertation). Retrieved from http://purl.galileo.usg.edu/uga%5Fetd/acar%5Fselcuk%5F201312%5Fphd. (Internet LXC16 2013 Acar, S.)

Adams, R. J. & Khoo, S. T. (1996). *Quest*. Melbourne, Australia: Australian Council for Educational Research.

Aliotti, N. C. & Blanton, W. E. (1969, February). *Some dimensions of creative thinking ability achievement, and intelligence in first grade*. Paper presented at the American Educational Research Association, Los Angeles, CA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581-594.

Antunes, A. P. & Almeida, L. (2007). Avaliar a criatividad: Contibutos para validade de alguns subtestes do TPCT (Torrance Pensamento Creativity Test). *Revista Psicologia e Educaçao, 6*(1), 37–53.

Ariffin, S. R., Katran, F. A., Badib, A. A. N., & Rashid, N. A. (2011). Validity and reliability of the Malaysian creativity and innovation instrument (MyCrIn) using the Rasch

measurement model. *Recent Researches in E-Activities*, 59-64. Retrieved from http://www.wseas.us/e-library/conferences/2011/Jakarta/EACT/EACT-08.pdf

Aslan, A. E. (2001). Torrance yaratıcı düşünce testi'nin Türkçe versiyonu [Turkish version of the Torrance tests of creative thinking]. *Marmara Üniversitesi Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi, 14*, 19-40.

Aslan, A. E. & Puccio, G. J. (2006). Developing and testing a Turkish version of Torrance's tests of creative thinking: A study of adults. *Journal of Creative Behavior, 40*, 163-177.

Auzmendi, E., Villa, A., & Abedi, J. (1996). Reliability and validity of a newly constructed multiple-choice creativity instrument. *Creativity Research Journal, 9*, 89-95.

Awamleh, H., Al Farah, Y., & El-Zraigat, I. (2012). The level of creative abilities dimensions according to Torrance formal test (B) and their relationship with some variables (sex, age, GPA). *International Education Studies, 5*(6), 138-147.

Baker, F. B. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Ball, O. E. & Torrance, E. P. (1984). *Torrance tests of creative thinking: Streamlined scoring workbook: Figural A.* Bensenville, IL: Scholastic Testing Service.

Berger, R. M., Guilford, J. P., & Christensen, P. R. (1957). A factor-analytic study of planning abilities. *Psychological Monographs: General and Applied, 71*(6), 1-31.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Campos, A., Lopez, A., Gonzalez, M. A., & Perez-Fabello, M. J. (2000). Aspects of creativity affected by imaging capacity. *North American Journal of Psychology, 2*, 313-322.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36,* 648-662.

Carroll, J. B. (1941). A factor analysis of verbal abilities. *Psychometrika, 6,* 279-307.

Chase, C. I. (1985). Review of the Torrance tests of creative thinking. In J. V. Mitchell (Ed.), *The ninth mental measurement yearbook* (pp. 1631-1632). Lincoln, NE: Buros Institute of Mental Measurements.

Cheng, Y., Kim, K. H., & Hull, M. F. (2010). Comparisons of creative styles and personality types between American and Taiwanese college students and the relationship between creative potential and personality types. *Psychology of Aesthetics, Creativity, and the Arts, 4,* 103–112.

Cho, S. H., Nijenhuis, J. T., van Vianen, A. E. M., Kim, H. B., & Lee, K. H. (2010).The relationship between diverse components of intelligence and creativity. *Journal of Creative Behavior, 44,* 125-137.

Clapham, M. M. (1998). Structure of figural forms A and B of the Torrance tests of creative thinking. *Educational and Psychological Measurement, 58,* 275-283.

Clapham, M. M. (2004). The convergent validity of the Torrance tests of creative thinking and creativity interest inventories. *Educational and Psychological Measurement, 64,* 828-841.

Cohen, R. J. & Swerdlik, M. E. (1999). *Psychological testing and assessment: An introduction to tests and measurement* (4th ed.). Mountain View, CA: Mayfield.

Cramond, C., Matthews-Morgan, J., Bandalos, D., & Zuo, L. (2005). A report on the 40-year follow-up of the Torrance tests of creative thinking: Alive and well in the new millennium. *Gifted Child Quarterly, 49,* 283-291.

Crawford, C. B & Nirmal, B. (1976). A multivariate study of measures of creativity, achievement, motivation, and intelligence in secondary school students. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 8*, 189-201.

Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory.* Mason, OH: Cengage Learning.

Cropley, A. J. & Clapson, L. (1971). Long term test-retest reliability of creativity tests. *British Journal of Educational Psychology, 41*, 206-208.

Dalbec, E. (1966). *Creative development over a three-year period in a Catholic Liberal Arts College* (Unpublished master's thesis). University of Minnesota, Minneapolis, MN.

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*, 109–117.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Psychology Press.

Esquivel, G. B. & Lopez, E. (1988). Correlations among measures of cognitive ability, creativity, and academic achievement for gifted minority children. *Perceptual and Motor Skills, 67,* 395-398.

Ewing, M. T., Salzberger, T., & Rudolf, R. S. (2005). An alternative approach to assessing cross-cultural measurement equivalence in advertising research. *Journal of Advertising, 34*(1), 17-36.

Ferracuti, S., Cannoni, E., Burla, F., & Lazzari, R. (1999). Correlations for the Rorschach with the Torrance tests of creative thinking. *Perceptual and Motor Skills, 89*, 863-870.

Ferrando, M. (2004). *Creatividad e inteligencias multiples* [Creativity and multiple intelligences] (Tesina de licenciatura). Universidad de Murcia, Murcia, Spain.

Ferrando, M. (2006). *Creatividad e Inteligencia Emocional: Un Estudio Empırico en Alumnos Con Altas Habilidades* (Tesis Doctoral). Publicada en la Universidad de Murcia, Murcia, Spain.

Ferrando, M., Ferrándiz, C., Bermejo, M., Sánchez, C., Parra, J., & Prieto, M. (2007). Estructura interna y baremación del Test de Pensamiento Creativo de Torrance. *Psicothema, 19*, 489-496.

Franck, K. & Rosen, E. (1949). A projective test of masculinity-femininity. *Journal of Consulting Psychology, 13*, 247-256.

Ghaemi, H. (2012). Is Rasch model without drawback? A reanalysis of Rasch model limitations. *Modern Journal of Language Teaching Methods, 1*(2), 31–38.

Goldman, R. J. (1965). The Minnesota tests of creative thinking. *Educational Research, 7*(1), 3-14.

Grover, B. L. (1963). *Some effects and correlates of different types of practice used in studying a topic in ninth grade classrooms* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Guilford, J. P. (1959). Three faces of intellect. *American Psychologist, 14*, 469-479.

Guilford, J. P. (1962). Factors that aid and hinder creativity. *Teachers College Record, 63*, 380-392.

Hagender, H. (1967). *Influence of creative writing experiences on general creative development* (Unpublished master's thesis). University of Minnesota, Minneapolis, MN.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Instructional Topics in Educational Measurement, 12*(3), 38-47.

Harkey, N. J. (1982). The Franck drawing completion test: A tool for research in sex-role identification. *Journal of Personality Assessment, 46*(1), 32-43.

Hattie, J. (1980). Should creativity tests be administered under testlike conditions? An empirical study of three alternative conditions. *Journal of Educational Psychology*, *72*(1), 87-98.

Hattie, J., Jaeger, R. M. & Bond, L. (1999). Chapter 11: Persistent methodological questions in educational testing. *Review of Research in Education, 24,* 393-446.

Heausler, N. L. & Thompson, B. (1988). Structure of the Torrance tests of creative thinking. *Educational and Psychological Measurement, 48*, 463-468.

Hunsaker, S. L., Abeel, L. B., & Callahan, C. M. (1991, June). *Instrument use in the identification of gifted and talented children*. Paper presented at the Meeting of the Jacob K. Javits Gifted and Talented Education Program Grant Recipients, Washington, DC.

Johnson, D. M. & Reynolds, F. (1941). A factor analysis of verbal ability. *The Psychological Record, 4*, 183-195.

Johnson, L. D. (1985). Creative thinking potential: Another example of u-shaped development? *Creative Child and Adult Quarterly, 10,* 146-159.

Johnson, R. A. (1974). Differential effects of reward versus no-reward instructions on the creative thinking of two economic levels of elementary school children. *Journal of Educational Psychology 66*, 530-633.

Karwowski, M. (2014). Creative mindsets: Measurement, correlates, consequences. *Psychology of Aesthetics, Creativity, and the Arts, 8*, 62–70.

Kaufman, J. C., Plucker, J. A., & Russell, C. M. (2012). Identifying and assessing creativity as a component of giftedness. *Journal of Psychoeducational Assessment, 30*, 60-73.

Kim, K. H. (2006). Is creativity unidimensional or multidimensional? Analyses of the Torrance tests of creative thinking. *Creativity Research Journal, 18*, 251–259.

Kim, K. H., Cramond, B., & Bandalos, D. L. (2006). The latent structure and measurement invariance of scores on the Torrance tests of creative thinking-figural. *Educational and Psychological Measurement, 66*, 459-477.

Kim, K. H. & VanTassel-Baska, J. (2010). The relationship between creativity and behavior problems among underachieving elementary and high school students. *Creativity Research Journal, 22*, 185–193.

Krumm, G., & Lemos, V. (2011). Análisis de las propiedades psicométricas de la prueba de figuras del test de pensamiento creativo de Torrance (TTCT). Forma B, en la provincia de Entre Ríos, Argentina [Analysis of the psychometric properties of the Figural Torrance Test of Creative Thinking (TTCT) Form B in the province of Entre Rios, Argentina]. In M. C. Richaud de Minzi & V. Lemos (Eds.), *Psicología y otras ciencias del comportamiento. Compendio de investigaciones actuales* (pp. 731–748). Libertador San Martín, Entre Ríos: Universidad Adventista del Plata.

Krumm, G., Lemos, V., & Filippetti, V. A. (2014). Factor structure of the Torrance tests of creative thinking figural form b in Spanish-speaking children: Measurement invariance across gender. *Creativity Research Journal, 26*, 72–81.

Linacre, J. M. (1999). Category disordering (disordered categories) vs. threshold disordering (disordered thresholds). *Rasch Measurement Transactions, 13,* 675. Retrieved from http://www.rasch.org/rmt/rmt131a.htm

Linacre, J. M. (2002a). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions, 16, 878. Retrieved from http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2002b). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre J. M. (2010). Transitional categories and usefully disordered thresholds. *Online Educational Research Journal, 1*(1), 1–10. Retrieved from http://www.oerj.org/View?action=viewPDF&paper=2

Linacre, J. M. (2013). *A user's guide to Facets Rasch-model computer programs*. Chicago, IL: MESA Press.

Linacre, J. M. (2013). *Facets Rasch measurement computer program* (Version 3.71.3). Chicago, IL: MESA Press.

Linacre, J. M. (2016). *A user's guide to Winsteps Rasch-model computer programs*. Chicago, IL: MESA Press.

López, O. (2001). *Evaluación y desarrollo de la creatividad* [Evaluation and development of creativity]. Doctoral thesis, Universidad de Murcia, Spain.

Mackler, B. (1962). *Creativity and life style* (Unpublished doctoral dissertation). University of Kansas, Lawrence, KS.

Matud, M. P. & Grande, C. R. J. (2007). Gender differences in creative thinking. *Personality and Individual Differences, 43,* 1137–1147.

Meade, A. W. & Lautenschlager, G. J. (2004, April). *Same question different answers: CFA and two IRT approaches to measurement invariance*. Symposium presented at the 19th Annual Conference of Society for Industrial and Organizational Psychology, Chicago, IL. Retrieved from http://www4.ncsu.edu/~awmeade/Links/Papers/CFA&IRT_MEI%28SIOP04%29.pdf

Mednick, S. A. (1968). The remote associates test. *The Journal of Creative Behavior, 2,* 213–214.

Nakano, T. C. & Primi, R. (2014). Rasch-Master's partial credit model in the assessment of children's creativity in drawings. *Spanish Journal of Psychology, 17,* 1-16.

Nakano T. C., Wechsler S. M., & Primi R. (2011). *Teste de Criatividade Figural Infantil* [Test of the Children's Figural Creativity]. São Paulo, Brasil: Editora Vetor.

Ogletree, E. J. & Ujlaki, W. (1973). Effects of social class status on tests of creative behavior. *The Journal of Educational Research, 67*(4), 149-152.

Oliveira, E. P. L. (2007). *Alunos sobredotados: A aceleraçao escolar como resposta educativa* (Tese de doutoramento). Universidade do Minho, Braga, Portugal.

Palaniappan, A. K. (2008). Influence of intelligence on the relationship between creativity and academic achievement: A comparative study. *The International Journal of Learning, 15*(7), 267-277.

Palaniappan, A. K. & Torrance, E. P. (2001). Comparison between regular and streamlined versions of scoring of Torrance tests of creative thinking. *The Korean Journal of Thinking & Problem Solving, 11*(2), 5-7.

Pelton, T. (2002, April). *Where are the limits to the Rasch advantage?* Paper presented to the International Objective Measurement Workshop (IOMW), New Orleans, LA.

Plucker, J. A. & Makel, M. C. (2010). Assessment of creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity (pp. 48-73)*. New York, NY: Cambridge University Press.

Prieto, M. D., López, O., Ferrándiz, C., & Bermejo, M. R. (2003). Adaptación de la prueba figurativa del Test de Pensamiento Creativo de Torrance en una muestra de los primeros niveles educativos [Adaptation of the figurative test of Creative Thinking Torrance Test in a sample of the first educational levels]. *Revista de Investigación Educativa, 21*, 201–213.

Prieto, M. D., Parra, J., Ferrándo, M., Ferrándiz, C., Bermejo, M. R., & Sánchez, C. (2006). Creative abilities in early childhood. *Journal of Early Childhood Research, 4*, 277-290.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Rampaul, W. E. Singh, M., & Didyk, J. (1984). The relationship between academic achievement, self-concept, creativity, and teacher expectations among native children in a Northern Manitoba School. *The Alberta Journal of Educational Research, 30*, 213-225.

Reise, S. P., Keith, F. W., & Robin, H. P. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Rudowicz, E., Lok, D., & Kitto, J. (1995). Use of the Torrance tests of creative thinking in an exploratory study of creativity in Hong Kong primary school children: A cross-cultural comparison. *International Journal of Psychology, 30*, 417-430.

Runco, M. A. (1986). Flexibility and originality in children's divergent thinking. *Journal of Psychology, 120*, 345-352.

Runco, M. A. (2001). Development and psychometric integrity of a measure of ideational behavior. *Creativity Research Journal, 13*, 393-400.

Runco, M. A., Millar, G. Acar, A., & Cramond, B. (2010). Torrance tests of creative thinking as predictors of personal and public achievement: A fifty-year follow-up. *Creativity Research Journal, 22*, 361–368.

Saeki, N., Fan, X., & Van Dusen, L. V. (2001). A comparative study of creative thinking of American and Japanese college students. *Journal of Creative Behavior, 35,* 24-38.

Schoenfeld, N. (1941). The metaphor of 'closure'. *Psychological Review, 48*, 487-497.

Sick, J. (2011). Rasch measurement in language education part 6: Rasch measurement and factor analysis. *SHIKEN: JALT Testing and Evaluation SIG Newsletter, 15*(1), 15-17. Retrieved from jalt.org/test/PDF/Sick6.pdf

Sikka, A. (1992, November). *Responses of African-American students on the Torrance tests of creative thinking (figural)*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Knoxville, TN.

Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal, 18*, 87-98.

Tannehill, R. L. (1992, November). *Assessing creativity in Native American students using the Torrance tests of creative thinking, figural form A.* Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Knoxville, TN.

Taylor, C. W. (1947). A factorial study of fluency in writing. *Psychometrika, 12*(4), 239-262

Taylor, C. W. (1988). Various approaches to and definitions of creativity. In R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives* (pp. 99-121). Cambridge, England: Cambridge University Press.

Thurstone, L. L. (1938). *Primary mental abilities*. Psychometric Monographs, No. 1. Chicago, IL: University of Chicago Press.

Torrance, E. P. (1962). *Guiding creative talent.* Englewood Cliffs, NJ: Prentice-Hall.

Torrance, E. P. (1966). *Torrance tests of creative thinking: Norms technical manual (research edition).* Princeton, NJ: Personnel Press, Inc.

Torrance, E. P. (1968). *Minnesota studies of creative behavior: 1958-1966*. Greensboro, NC: The Creativity Research Institute of the Richardson Foundation.

Torrance, E. P. (1972). Creative young women in today's world. *Exceptional Children, 38*, 597-603.

Torrance, E. P. (1974). *Torrance tests of creative thinking: Norms technical manual, figural forms A & B.* Bensenville, IL: Scholastic Testing Service.

Torrance, E. P. (1979). *The search for satori and creativity.* Buffalo, NY: Creative Education Foundation.

Torrance, E. P. (1981). Empirical validation of criterion-referenced indicators of creative ability through a longitudinal study. *Creative Child and Adult Quarterly, 6*, 136-140.

Torrance, E. P. (1984). *Torrance test of creative thinking: Norms technical manual, figural (streamlined) forms A & B*. Bensenville, IL: Scholastic Testing Service.

Torrance, E. P. (1988). The nature of creativity as manifest in its testing. In R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives (pp. 43-75).* New York, NY: Cambridge University Press.

Torrance, E. P. (1990). *Torrance test of creative thinking: Norms technical manual, figural (streamlined) forms A & B*. Bensenville, IL: Scholastic Testing Service.

Torrance, E. P. (1998). *Torrance test of creative thinking: Norms technical manual, figural (streamlined) forms A & B*. Bensenville, IL: Scholastic Testing Service.

Torrance, E. P. (2000). *Research review for the Torrance tests of creative thinking figural and verbal forms A and B*. Bensenville, IL: Scholastic Testing Service.

Torrance, E. P. (2006a). *Thinking creatively with pictures: Figural Booklet A.* Bensenville, IL: Scholastic Testing Service.

Torrance, E. P. (2006b). *Thinking creatively with pictures: Figural Booklet B.* Bensenville, IL: Scholastic Testing Service.

Torrance, E. P. (2008). *Torrance test of creative thinking: Norms technical manual, figural (streamlined) forms A & B*. Bensenville, IL: Scholastic Testing Service.

Torrance, E. P., Ball, O. E., & Safter, H. T. (1992). *Torrance tests of creative thinking: Streamlined scoring guide for figural forms A & B*. Bensenville, IL: Scholastic Testing Service.

Torrance, E. P. & Safter, H. T. (1999). *Making the creative leap beyond*. Buffalo, NY: Creative Education Foundation Press.

Tran, N. (2004). Cultural dimensions in creativity: A preliminary study about creativity among the Vietnamese people in America. *Education and Society*, *22*(2), 71-81.

Voss, D. H. (1997). *Determining test fairness and differential validity of scores for the Torrance tests of creative thinking for kindergarten students* (Unpublished doctoral dissertation). Texas Tech University, Lubbock, TX.

Wang, C., Ho, H., Cheng, C., & Cheng, Y. (2014). Application of the Rasch model to the measurement of creativity: The creative achievement questionnaire. *Creativity Research Journal, 26*, 62–71.

Ward, W. C. (1969). Rate and uniqueness in children's creative responding. *Child Development, 40,* 869-878.

Wechsler, S. M., Vendramini, C. M. M., & Oakland, T. (2012). Thinking and creative styles: A validity study. *Creativity Research Journal, 24,* 235–242.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychological Bulletin, 50*(5), 362-370.

Wilson, R. C., Guilford, J. P., Christensen, P. R., & Lewis, D. J. A. (1954). A factor analytic study of creative-thinking abilities. *Psychometrika, 19,* 297-311.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 1994, 8,* 370. Retrieved from http://www.rasch.org/rmt/rmt83b.htm

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA.

Wright, B. D. & Masters, G. N. (2002). Number of person or item strata: (4*Separation + 1)/3. *Rasch Measurement Transactions, 16,* 888. Retrieved from http://www.rasch.org/rmt/rmt163f.htm

Yong, L. M. S. (1994). Relations between creativity and intelligence among Malaysian pupils. *Perceptual and Motor Skills, 79,* 739-742.

Table 1

Mean item scores for all the variables scored at the item level

	Items																
Variables	A1 I1	A2 I1	A2 I2	A2 I3	A2 I4	A2 I5	A2 I6	A2 I7	A2 I8	A2 I9	A2 I10	A3 I1	A3 I2	A3 I3	A3 I4	A3 I5	A3 I6
Fluency	0.94	0.99	0.82	0.90	0.85	0.92	0.84	0.85	0.82	0.94	0.87	0.92	0.92	0.85	0.91	0.84	0.77
Originality	0.83	0.63	0.47	0.84	0.65	0.72	0.58	0.79	0.56	0.57	0.46	0.34	0.44	0.48	0.43	0.48	0.39
Elaboration-I	NA	1.51	1.37	1.30	1.64	1.51	1.22	1.54	1.31	1.17	1.04	1.56	1.93	1.64	1.68	1.64	1.45
Abs. of Titles	1.10	0.38	0.37	0.36	0.35	0.34	0.33	0.30	0.28	0.45	0.32	NA	NA	NA	NA	NA	NA
Closure	N.A	0.96	1.14	1.11	1.43	0.64	1.36	0.98	0.93	0.90	0.92	NA	NA	NA	NA	NA	NA
Emotional	0.23	0.12	0.02	0.03	0.06	0.15	0.05	0.02	0.04	0.06	0.02	0.04	0.04	0.05	0.02	0.02	0.05
Storytelling	0.47	0.08	0.04	0.12	0.08	0.06	0.07	0.07	0.04	0.05	0.01	0.06	0.06	0.10	0.10	0.05	0.06
Movement	0.40	0.21	0.06	0.14	0.17	0.02	0.11	0.10	0.08	0.16	0.02	0.07	0.06	0.06	0.08	0.06	0.03
Ex. of Titles	0.43	0.18	0.09	0.12	0.14	0.13	0.11	0.09	0.09	0.13	0.08	0.12	0.10	0.09	0.09	0.07	0.07
Synthesis 2	NA	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	NA	NA	NA	NA	NA	NA
Synthesis 3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.01	0.03	0.03	0.02	0.02	0.01
Unusual	0.70	0.12	0.14	0.49	0.37	0.24	0.41	0.32	0.31	0.15	0.10	0.14	0.14	0.20	0.20	0.21	0.21
Internal	0.11	0.06	0.05	0.02	0.04	0.07	0.04	0.05	0.03	0.04	0.01	0.06	0.08	0.06	0.07	0.08	0.06
Extending	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.46	0.48	0.50	0.30	0.31	0.29
Humor	0.12	0.07	0.02	0.04	0.05	0.05	0.04	0.02	0.02	0.04	0.02	0.03	0.01	0.03	0.02	0.03	0.04
Richness	0.52	0.14	0.14	0.21	0.24	0.15	0.17	0.16	0.18	0.12	0.10	0.07	0.13	0.14	0.10	0.10	0.07
Colorfulness	0.54	0.10	0.09	0.15	0.32	0.11	0.15	0.11	0.10	0.13	0.05	0.04	0.09	0.15	0.08	0.08	0.11
Fantasy	0.34	0.08	0.05	0.05	0.12	0.05	0.05	0.07	0.05	0.09	0.04	0.03	0.04	0.06	0.06	0.07	0.05

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. NA implies that the item was not scored for this variable.

Table 2

Rasch item parameters for fluency

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.29*	1.10	-0.81	-0.57	-1.22	0.64	0.84	.40
A2 I1	1.05	0.20	-2.72	-2.14	-3.55	1.41	0.78	.43
A2 I2	0.99	0.00	0.71	0.76	0.63	0.13	0.27	.78
A2 I3	1.08	0.40	-0.14	-0.57	0.29	-0.86	-1.48	.14
A2 I4	0.90	-0.70	0.51	0.62	0.35	0.27	0.54	.59
A2 I5	1.11	0.50	-0.51	-0.40	-0.67	0.27	0.42	.67
A2 I6	0.95	-0.30	0.53	0.65	0.36	0.29	0.58	.56
A2 I7	0.93	-0.40	0.48	0.72	0.13	0.58	1.14	.25
A2 I8	1.11	0.90	0.72	-0.36	1.11	-0.76	-1.58	.11
A2 I9	0.97	0.00	-0.83	-1.28	-0.38	-0.90	-1.24	.21
A2 I10	1.06	0.40	0.22	0.11	0.34	-0.24	-0.44	.65
A3 I1	0.98	0.00	-0.34	-0.18	-0.62	0.44	0.65	.51
A3 I2	0.98	0.00	-0.32	-0.29	-0.34	0.05	0.07	.94
A3 I3	0.86	-0.90	0.54	0.52	0.57	-0.05	-0.09	.92
A3 I4	0.96	-0.10	-0.10	-0.21	0.05	-0.26	-0.42	.67
A3 I5	0.85	-1.10	0.69	0.75	0.62	0.13	0.24	.80
A3 I6	1.01	0.10	1.36	1.36	1.38	-0.02	-0.04	.96

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty. * shows the misfitting item.

Table 3

Rasch item parameters for originality

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.07	0.50	-1.41	-1.32	-1.56	0.24	0.53	.59
A2 I1	1.11	1.70	-0.23	0.00	-0.58	0.58	1.68	.09
A2 I2	1.04	0.70	0.51	0.31	0.80	-0.49	-1.44	.15
A2 I3	1.02	0.10	-1.47	-1.61	-1.32	-0.29	-0.66	.50
A2 I4	0.97	-0.40	-0.36	-0.43	-0.27	-0.16	-0.45	.65
A2 I5	1.01	0.10	-0.71	-0.64	-0.80	0.16	0.44	.66
A2 I6	1.03	0.50	-0.02	-0.03	-0.01	-0.02	-0.06	.95
A2 I7	0.96	-0.30	-1.16	-1.04	-1.32	0.28	0.69	.48
A2 I8	1.01	0.10	0.06	-0.15	0.34	-0.48	-1.42	.15
A2 I9	0.98	-0.30	0.02	-0.21	0.34	-0.55	-1.62	.10
A2 I10	0.97	-0.60	0.54	0.64	0.40	0.24	0.72	.47
A3 I1	0.99	-0.10	1.13	1.22	0.98	0.25	0.67	.50
A3 I2	1.03	0.50	0.62	0.75	0.43	0.32	0.92	.36
A3 I3	0.96	-0.60	0.46	0.50	0.41	0.08	0.24	.80
A3 I4	0.97	-0.50	0.68	0.57	0.87	-0.30	-0.83	.40
A3 I5	0.91	-1.80	0.45	0.59	0.24	0.35	0.96	.33
A3 I6	0.94	-1.00	0.88	0.87	0.92	-0.05	-0.14	.89

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty.

Table 4

Rasch parameters for elaboration

Activities	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
1	0.81	-1.40	-0.83	-0.69	-1.00	0.31	1.23	.22
2	0.93	-0.40	0.86	0.99	0.71	0.28	0.90	.37
3	1.20	1.40	-0.04	0.09	-0.10	0.19	1.78	.08

Note. δ is average item (i.e., activity) difficulty.

Table 5

Rasch item parameters for elaboration-I

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A2 I1	1.09	0.70	-0.03	-0.01	-0.06	0.05	0.43	.66
A2 I2	1.03	0.30	0.04	0.10	-0.04	0.14	1.27	.20
A2 I3	0.91	-0.60	0.08	0.06	0.10	-0.03	-0.29	.77
A2 I4	0.93	-0.50	-0.09	-0.04	-0.17	0.13	1.18	.23
A2 I5	1.20	1.60	-0.03	-0.07	0.02	-0.08	-0.75	.45
A2 I6	1.07	0.50	0.12	0.11	0.13	-0.02	-0.17	.86
A2 I7	1.05	0.40	-0.05	0.02	-0.14	0.16	1.47	.14
A2 I8	0.86	-1.00	0.08	0.08	0.07	0.01	0.11	.91
A2 I9	0.98	-0.10	0.17	0.10	0.27	-0.17	-1.30	.19
A2 I10	1.06	0.40	0.25	0.20	0.31	-0.11	-0.83	.40
A3 I1	1.08	0.70	-0.05	0.01	-0.12	0.13	1.13	.26
A3 I2	0.75*	-2.40*	-0.22	-0.22	-0.20	-0.02	-0.21	.83
A3 I3	1.05	0.40	-0.08	-0.10	-0.05	-0.05	-0.41	.68
A3 I4	1.02	0.20	-0.11	-0.10	-0.12	0.02	0.16	.87
A3 I5	1.02	0.10	-0.07	-0.11	-0.01	-0.10	-0.81	.41
A3 I6	1.06	0.50	0.01	-0.06	0.13	-0.19	-1.44	.15

Note. The letter A in A2 and A3 refers to *activity*, as in Activity 2. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is average item difficulty. * shows the misfitting item.

Table 6

Rasch item parameters for abstractness of titles

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.08	0.80	-1.89	-1.84	-1.97	0.12	0.47	.63
A2 I1	0.85	-0.90	0.08	0.05	0.12	-0.07	-0.23	.81
A2 I2	1.03	0.20	0.14	-0.17	0.82	-0.98	-2.73	.01♦
A2 I3	1.02	0.10	0.17	0.08	0.33	-0.25	-0.76	.44
A2 I4	0.86	-0.80	0.21	0.19	0.24	-0.05	-0.14	.89
A2 I5	1.11	0.70	0.23	0.09	0.48	-0.39	-1.13	.25
A2 I6	0.94	-0.30	0.24	0.41	-0.03	0.44	1.34	.18
A2 I7	0.85	-0.90	0.38	0.61	0.06	0.55	1.61	.10
A2 I8	1.20	1.10	0.41	0.47	0.32	0.15	0.42	.67
A2 I9	0.79*	-1.50	-0.21	0.07	-0.60	0.68	2.32	.02♦
A2 I10	1.31*	1.70	0.25	0.08	0.56	-0.48	-1.34	.18

Note. The letter A in A1 and A2 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is average item difficulty. * shows the misfitting items. ♦ shows the items with differential item functioning.

Table 7

Rasch item parameters for resistance to premature closure

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
1	0.97	-0.30	0.15	0.29	-0.04	0.32	1.60	.11
2	1.03	0.30	-0.15	-0.19	-0.08	-0.11	-0.51	.60
3	0.98	-0.20	-0.12	-0.10	-0.13	0.03	0.15	.88
4	1.19	1.70	-0.73	-0.67	-0.80	0.13	0.57	.56
5	1.09	1.00	0.71	0.53	0.96	-0.42	-1.93	.06
6	1.08	0.80	-0.58	-0.66	-0.47	-0.19	-0.86	.39
7	0.97	-0.30	0.10	0.31	-0.19	0.50	2.41	.02♦
8	0.99	-0.10	0.18	0.08	0.30	-0.22	-1.06	.29
9	0.91	-1.20	0.25	0.19	0.32	-0.14	-0.67	.50
10	0.87	-1.70	0.20	0.21	0.18	0.03	0.14	.88

Note. δ is average item difficulty. ♦ shows the item with differential item functioning.

Table 8

Rasch item parameters for emotional expressiveness

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.18	2.0	-2.59	-2.83	-2.19	-0.64	-1.16	.25
A2 I1	1.13	0.80	-1.41	-1.23	-1.73	0.50	0.84	.40
A2 I2	0.98	0.10	0.92	1.34	0.27	1.06	0.97	.33
A2 I3	1.17	0.50	0.64	0.49	1.11	-0.62	-0.51	.61
A2 I4	0.87	-0.50	-0.40	0.21	-1.26	1.47	2.02	.04♦
A2 I5	1.21*	1.50	-1.73	-2.17	-0.66	-1.51	-2.22	.03♦
A2 I6	0.77*	-0.80	-0.20	-0.39	0.27	-0.66	-0.72	.47
A2 I7	0.73*	-0.50	0.83	0.72	1.11	-0.39	-0.31	.75
A2 I8	0.91	-0.10	0.12	0.05	0.27	-0.23	-0.24	.81
A2 I9	1.01	0.10	-0.29	-0.08	-0.66	0.58	0.75	.45
A2 I10	1.08	0.30	0.82	1.20	0.27	0.92	0.84	.40
A3 I1	0.84	-0.40	0.40	0.46	0.30	0.16	0.16	.87
A3 I2	0.93	-0.10	0.40	0.46	0.30	0.16	0.16	.87
A3 I3	0.87	-0.40	0.04	0.20	-0.25	0.44	0.52	.60
A3 I4	1.18	0.50	1.22	2.06	0.30	1.76	1.34	.18
A3 I5	0.82	-0.20	1.12	0.79	1.86	-1.07	-0.59	.55
A3 I6	1.06	0.30	0.10	0.15	-0.02	0.17	0.18	.85

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty. * shows the misfitting items. ♦ shows the items with differential item functioning.

Table 9

Rasch item parameters for storytelling articulateness

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.22*	1.90	-3.95	-3.85	-4.12	0.27	0.51	.61
A2 I1	1.03	0.20	-0.19	-0.63	0.74	-1.37	-1.81	.07
A2 I2	1.13	0.50	0.75	0.76	0.74	0.02	0.03	.97
A2 I3	0.95	-0.20	-0.79	-0.74	-0.85	0.12	0.21	.83
A2 I4	1.03	0.20	-0.07	-0.30	0.36	-0.66	-0.93	.35
A2 I5	1.04	0.20	0.16	0.48	-0.22	0.70	0.99	.32
A2 I6	0.85	-0.50	0.16	0.25	0.05	0.20	0.28	.77
A2 I7	1.04	0.20	0.15	1.07	-0.66	1.74	2.27	.03♦
A2 I8	1.03	0.10	0.73	1.07	0.36	0.72	0.84	.40
A2 I9	1.07	0.30	0.42	0.46	0.36	0.11	0.14	.88
A2 I10	1.06	0.30	2.17	2.27	2.05	0.23	0.15	.87
A3 I1	1.01	0.10	0.30	0.43	0.10	0.33	0.43	.67
A3 I2	0.84	-0.50	0.29	-0.02	1.02	-1.04	-1.15	.25
A3 I3	0.71*	-1.60	-0.41	-0.66	0.10	-0.76	-1.08	.28
A3 I4	0.98	0.00	-0.46	-0.55	-0.29	-0.26	-0.38	.70
A3 I5	1.06	0.20	0.54	.35	0.93	-0.58	-0.61	.54
A3 I6	0.93	-0.10	0.20	.27	0.07	0.20	0.24	.80

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty. * shows the misfitting items. ♦ shows the item with differential item functioning.

Table 10

Rasch item parameters for movement or action

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.10	1.03	-2.53	-2.37	-2.81	0.44	1.07	.28
A2 I1	1.10	0.90	-1.26	-1.28	-1.23	-0.05	-0.11	.90
A2 I2	1.06	0.20	0.49	0.16	1.02	-0.87	-1.16	.24
A2 I3	0.94	-0.40	-0.65	-0.55	-0.77	0.21	0.44	.66
A2 I4	1.08	0.60	-0.89	-0.77	-1.03	0.27	0.57	.56
A2 I5	1.01	0.10	1.82	2.28	1.47	0.81	0.64	.52
A2 I6	0.88	-0.60	-0.29	-0.06	-0.54	0.49	0.88	.37
A2 I7	0.94	-0.20	-0.13	0.49	-0.63	1.12	1.85	.07
A2 I8	1.06	0.30	0.03	0.26	-0.19	0.45	0.74	.45
A2 I9	0.98	-0.10	-0.85	-1.26	-0.19	-1.06	-2.11	.04♦
A2 I10	1.05	0.20	1.79	2.22	1.47	0.76	0.60	.54
A3 I1	0.92	-0.20	0.19	-0.26	1.21	-1.47	-1.76	.08
A3 I2	1.12	0.50	0.30	-0.13	1.21	-1.34	-1.58	.11
A3 I3	0.86	-0.40	0.45	0.49	0.40	0.08	0.11	.90
A3 I4	0.89	-0.40	0.05	-0.28	0.71	-0.99	-1.35	.17
A3 I5	0.81	-0.60	0.34	0.38	0.28	0.10	0.13	.89
A3 I6	1.08	0.30	1.15	2.08	0.46	1.62	1.35	.17

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty. ♦ shows the item with differential item functioning.

Table 11

Rasch item parameters for expressiveness of titles

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.35*	3.60*	-3.15	-2.97	-3.46	0.49	0.98	.32
A2 I1	0.74*	-1.80	-0.86	-0.94	-0.75	-0.19	-0.35	.72
A2 I2	1.31*	1.30	0.45	0.02	1.37	-1.36	-1.71	.09
A2 I3	0.84	-0.80	-0.07	-.45	0.62	-1.06	-1.59	.11
A2 I4	0.81	-1.10	-0.35	-0.58	0.04	-0.63	-1.02	.31
A2 I5	0.93	-0.30	-0.16	-.058	0.62	-1.20	-1.80	.07
A2 I6	0.95	-0.10	-0.01	0.01	-0.04	0.05	0.08	.93
A2 I7	0.82	-0.80	0.34	1.01	-0.37	1.38	2.04	.04♦
A2 I8	1.12	0.60	0.30	0.16	0.52	-0.36	-0.52	.60
A2 I9	1.00	0.00	-0.31	0.14	-0.86	1.00	1.68	.09
A2 I10	1.06	0.30	0.53	0.34	0.87	-0.53	-0.70	.48
A3 I1	1.06	0.30	0.07	-0.51	1.34	-1.85	-2.35	.02♦
A3 I2	0.92	-0.30	0.40	0.69	0.02	0.67	0.98	.32
A3 I3	1.08	0.40	0.53	1.47	-0.47	1.94	2.72	.01♦
A3 I4	0.94	-0.10	0.61	0.67	0.54	0.12	0.16	.86
A3 I5	0.80	-0.70	0.90	1.38	0.24	1.15	1.44	.15
A3 I6	1.04	0.20	0.78	0.97	0.49	0.47	0.58	.56

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty. * shows the misfitting items. ♦ shows the items with differential item functioning.

Table 12

Rasch item parameters for unusual visualization

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	0.97	-0.30	-2.27	-2.17	-2.44	0.26	0.69	.49
A2 I1	1.03	0.20	0.95	1.10	0.72	0.38	0.77	.44
A2 I2	1.06	0.40	0.73	0.73	0.72	0.00	0.01	.99
A2 I3	1.00	0.00	-1.26	-1.53	-0.88	-0.64	-1.88	.06
A2 I4	1.01	0.10	-0.68	-0.83	-0.45	-0.38	-1.09	.27
A2 I5	1.07	0.70	0.01	-0.19	0.36	-0.55	-1.36	.17
A2 I6	0.94	-1.00	-0.86	-0.64	-1.19	0.55	1.60	.11
A2 I7	0.97	-0.40	-0.44	-.033	-0.60	0.27	0.77	.44
A2 I8	0.91	-1.30	-0.36	-0.23	-0.53	0.30	0.82	.41
A2 I9	1.04	0.30	0.66	0.89	0.35	0.53	1.18	.23
A2 I10	0.98	0.00	1.11	0.86	1.66	-0.79	-1.31	.19
A3 I1	0.93	-0.30	0.72	0.82	0.56	0.26	0.53	.59
A3 I2	1.03	0.20	0.71	0.90	0.42	0.48	1.00	.31
A3 I3	0.98	-0.10	0.28	0.14	0.56	-0.42	-0.91	.36
A3 I4	0.98	-0.10	0.25	0.03	0.70	-0.67	-1.42	.15
A3 I5	0.99	0.00	0.22	0.35	0.01	0.35	0.80	.42
A3 I6	1.05	0.40	0.24	0.26	0.22	0.04	0.10	.92

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty.

Table 13

Rasch item parameters for internal visualization

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	0.95	-0.20	-0.93	-0.80	-1.07	0.27	0.48	.63
A2 I1	1.03	0.10	-0.21	-.039	0.02	-0.41	-0.59	.55
A2 I2	1.03	0.10	-0.09	1.56	-0.83	2.39	2.18	.03♦
A2 I3	1.04	0.20	1.10	0.84	1.48	-0.64	-0.51	.61
A2 I4	0.92	-0.10	0.18	0.40	-0.02	0.42	0.52	.60
A2 I5	0.93	-0.20	-0.47	-0.63	-0.28	-0.35	-0.54	.59
A2 I6	0.99	0.00	0.32	0.37	0.26	0.10	0.12	.90
A2 I7	0.96	0.00	-0.12	0.05	-0.28	0.33	0.45	.65
A2 I8	1.01	0.10	0.51	0.76	0.31	0.45	0.48	.63
A2 I9	1.05	0.20	0.35	0.84	-0.02	0.86	0.95	.34
A2 I10	0.98	0.30	2.22	1.53	2.22	-0.68	-0.39	.70
A3 I1	0.93	-0.10	-0.25	-0.04	-0.47	0.43	0.59	.55
A3 I2	0.98	0.00	-0.60	-0.88	-0.20	-0.69	-1.01	.31
A3 I3	1.01	0.10	-0.40	-0.71	0.09	-0.80	-1.07	.28
A3 I4	1.04	0.20	-0.57	-0.76	-0.30	-0.46	-0.65	.51
A3 I5	1.05	0.20	-0.66	-1.02	-0.06	-0.96	-1.30	.19
A3 I6	1.14	0.50	-0.35	-0.26	-0.47	0.22	0.28	.78

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty. ♦ shows the item with differential item functioning.

Table 14

Rasch item parameters for extending or breaking boundaries

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
1	1.05	0.60	-0.50	-0.46	-0.59	0.13	0.30	.76
2	1.06	0.60	-0.63	-0.72	-0.47	-0.25	-0.58	.56
3	0.96	-0.40	-0.71	-0.85	-0.47	-0.38	-0.88	.37
4	0.93	-0.70	0.59	0.94	0.07	0.87	1.91	.06
5	0.98	-0.10	0.57	0.82	0.18	0.63	1.37	.17
6	1.00	0.00	0.69	0.32	1.40	-1.08	-2.13	.03♦

Note. δ is item difficulty. ♦ shows the item with differential item functioning.

Table 15

Rasch item parameters for humor

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.45*	3.10*	-1.73	-1.24	-2.31	1.08	1.77	.08
A2 I1	0.94	-0.20	-0.92	-0.98	-0.82	-0.16	-0.22	.82
A2 I2	0.95	0.00	0.44	0.77	-0.01	0.78	0.74	.46
A2 I3	0.90	-0.20	-0.06	-0.38	0.74	-1.12	-0.97	.33
A2 I4	0.91	-0.20	-0.58	-0.38	-0.87	0.49	0.65	.51
A2 I5	0.95	-0.10	-0.57	-1.04	0.74	-1.78	-1.59	.11
A2 I6	0.88	-0.20	-0.12	-0.47	0.74	-1.21	-1.05	.29
A2 I7	0.91	0.00	0.32	0.11	0.74	-0.63	-0.52	.60
A2 I8	1.00	0.10	0.64	0.11	1.64	-1.53	-0.91	.36
A2 I9	1.21*	0.70	-0.28	0.21	-0.82	1.03	1.22	.22
A2 I10	0.98	0.10	0.63	0.57	0.74	-0.17	-0.14	.89
A3 I1	1.06	0.20	0.38	0.34	0.47	-0.13	-0.11	.91
A3 I2	1.08	0.30	1.11	1.51	0.47	1.04	0.71	.47
A3 I3	0.98	0.00	0.07	0.30	-0.34	0.64	0.65	.51
A3 I4	0.93	0.00	0.66	0.76	0.43	0.33	0.26	.79
A3 I5	0.87	-0.20	0.02	-0.07	0.29	-0.36	-0.30	.76
A3 I6	0.91	-0.10	-0.03	-0.10	0.20	-0.30	-0.26	.80

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty. * shows the misfitting items.

Table 16

Rasch item parameters for richness of imagery

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	0.90	-1.30	-2.38	-2.42	-2.33	-0.08	-0.21	.83
A2 I1	1.09	0.60	0.09	0.18	-0.04	0.21	0.44	.65
A2 I2	0.89	-0.60	0.15	0.40	-0.17	0.56	1.14	.25
A2 I3	0.91	-0.80	-0.52	-0.74	-0.17	-0.57	-1.29	.19
A2 I4	1.03	0.30	-0.73	-0.67	-0.82	0.16	0.39	.69
A2 I5	1.07	0.50	0.02	-0.29	0.59	-0.88	-1.69	.09
A2 I6	0.91	-0.60	-0.22	-0.34	-0.05	-0.28	-0.61	.54
A2 I7	0.95	-0.30	-0.08	0.01	-0.19	0.20	0.43	.66
A2 I8	0.97	-0.10	-0.29	0.00	-0.65	0.65	1.44	.15
A2 I9	1.00	0.00	0.30	0.04	0.77	-0.73	-1.30	.19
A2 I10	1.02	0.10	0.57	0.32	0.99	-0.67	-1.11	.26
A3 I1	1.06	0.30	0.94	1.54	0.31	1.24	1.79	.07
A3 I2	0.97	-0.10	0.17	0.31	-0.05	0.36	0.68	.49
A3 I3	0.88	-0.70	0.06	0.48	-0.49	0.97	1.89	.06
A3 I4	1.12	0.60	0.46	0.31	0.74	-0.42	-0.69	.49
A3 I5	1.08	0.40	0.50	0.38	0.72	-0.33	-0.54	.59
A3 I6	1.17	0.70	0.97	0.81	1.28	-0.48	-0.63	.52

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty.

Table 17

Rasch item parameters for colorfulness of imagery

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.05	0.60	-2.72	-2.87	-2.52	-0.35	-0.89	.37
A2 I1	1.09	0.50	0.27	0.36	0.18	0.17	0.32	.75
A2 I2	0.89	-0.40	0.44	0.10	1.00	-0.89	-1.41	.16
A2 I3	1.02	0.10	-0.24	-0.53	0.18	-0.71	-1.44	.15
A2 I4	1.05	0.60	-1.52	-1.18	-1.88	0.70	1.84	.07
A2 I5	1.08	0.40	0.20	0.21	0.18	0.03	0.06	.95
A2 I6	1.00	0.00	-0.26	-0.35	-0.14	-0.21	-0.43	.66
A2 I7	0.95	-0.20	0.10	0.05	0.18	-0.13	-0.25	.80
A2 I8	1.07	0.40	0.23	0.17	0.30	-0.13	-0.23	.81
A2 I9	1.04	0.20	-0.02	0.48	-0.47	0.95	1.84	.07
A2 I10	0.98	0.00	1.14	1.03	1.31	-0.28	-0.37	.71
A3 I1	1.01	0.10	1.20	1.29	1.09	0.20	0.25	.80
A3 I2	0.91	-0.30	0.38	0.16	0.76	-0.60	-0.92	.35
A3 I3	0.88	-0.80	-0.30	0.09	-0.76	0.85	1.73	.08
A3 I4	1.06	0.30	0.45	0.45	0.45	0.00	0.00	.99
A3 I5	0.89	-0.40	0.49	0.40	0.66	-0.27	-0.40	.69
A3 I6	0.99	0.00	0.15	0.03	0.34	-0.32	-0.51	.60

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty.

Table 18

Rasch item parameters for fantasy

Items	I-MNSQ	I-STZD	δ_{sample}	δ_{girls}	δ_{boys}	$\delta_{girls} - \delta_{boys}$	*t*	*p*
A1 I1	1.12	1.30	-3.13	-3.17	-3.08	-0.09	-0.18	.85
A2 I1	1.17	0.80	-0.29	-0.49	-0.04	-0.45	-0.67	.50
A2 I2	1.01	0.10	0.33	0.40	0.25	0.15	0.19	.84
A2 I3	0.96	0.00	0.48	0.40	0.60	-0.21	-0.25	.80
A2 I4	1.04	0.20	-0.93	-0.66	-1.20	0.54	0.96	.34
A2 I5	1.09	0.30	0.48	-0.09	1.84	-1.93	-1.70	.09
A2 I6	0.79*	-0.70	0.24	0.02	0.55	-0.53	-0.66	.50
A2 I7	0.92	-0.20	-0.02	0.01	-0.04	0.04	0.06	.95
A2 I8	0.91	-0.20	0.41	0.25	0.60	-0.35	-0.42	.67
A2 I9	1.22*	1.10	-0.41	-0.32	-0.50	0.18	0.28	.78
A2 I10	0.87	-0.30	0.58	0.25	1.08	-0.82	-0.89	.37
A3 I1	0.87	-0.20	0.99	1.05	0.91	0.14	0.14	.88
A3 I2	0.78*	-0.60	0.56	.67	0.42	0.25	0.29	.77
A3 I3	1.08	0.40	0.07	.36	-0.26	0.62	0.82	.41
A3 I4	0.95	-0.10	0.20	.36	-0.02	0.38	0.48	.63
A3 I5	1.13	0.50	-0.02	.64	-0.72	1.36	1.75	.08
A3 I6	0.79*	-0.60	0.45	0.29	0.74	-0.45	-0.48	.63

Note. The letter A in A1, A2, and A3 refers to *activity*, as in Activity 1. The letter I in I1 through I10 refers to *item*, as in Item 1. δ is item difficulty. * shows the misfitting items.

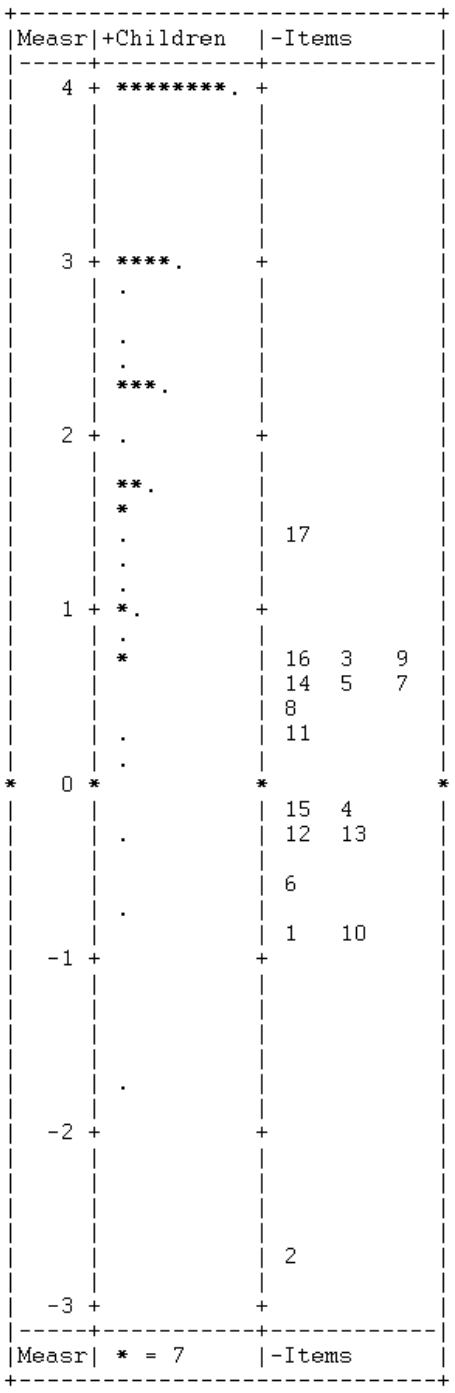

Figure 1. The variable map for fluency. In this figure, the first column represents the latent trait continuum from -3 to 4 logits. The second column shows the locations of the students on the latent trait continuum. * represents 7 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

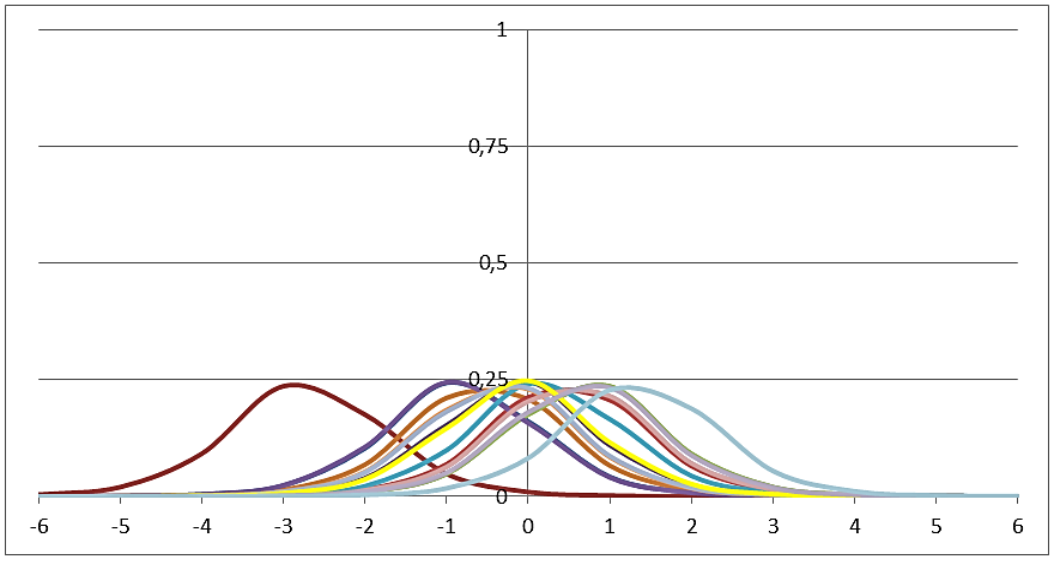

Figure 2. Item information functions for fluency. In this figure, each bell-shaped line represents the information function of one item. The horizontal axis represents the latent trait continuum from -6 to 6 logits. The vertical axis shows the amount of information provided by an item.

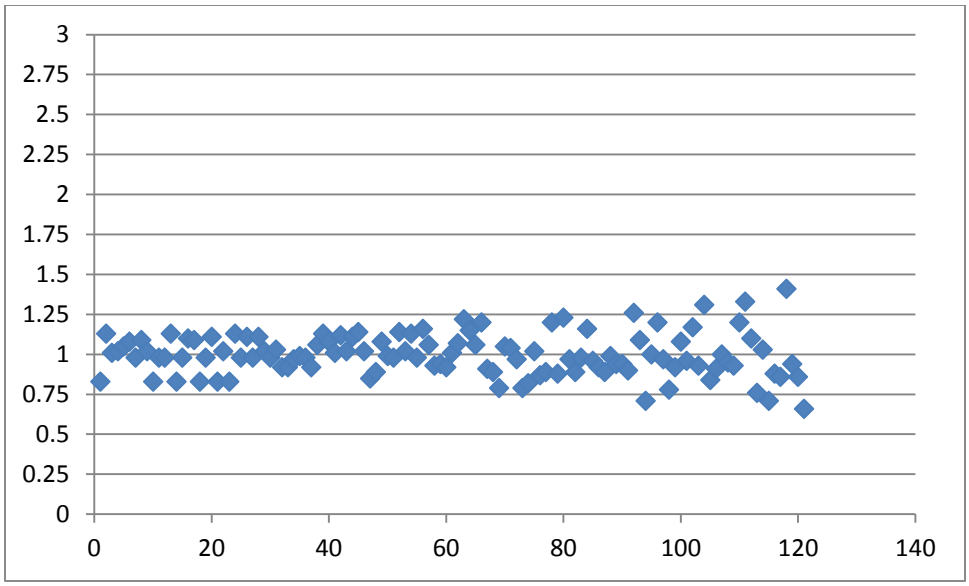

Figure 3. Person fit values for fluency. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

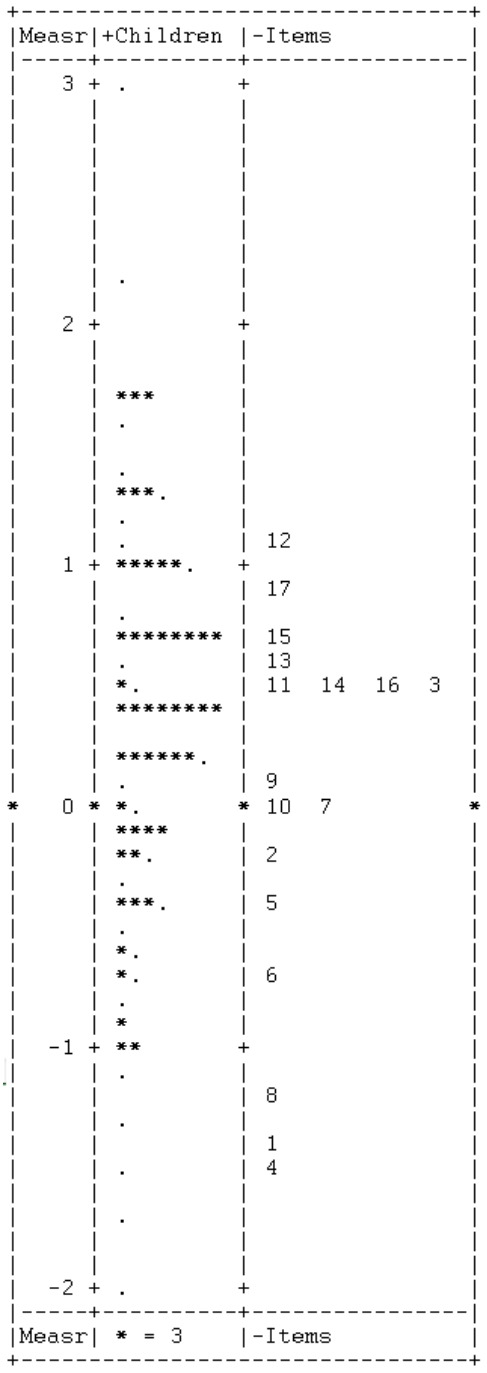

Figure 4. The variable map for originality. In this figure, the first column represents the latent trait continuum from -2 to 3 logits. The second column shows the locations of the students on the latent trait continuum. * represents 3 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

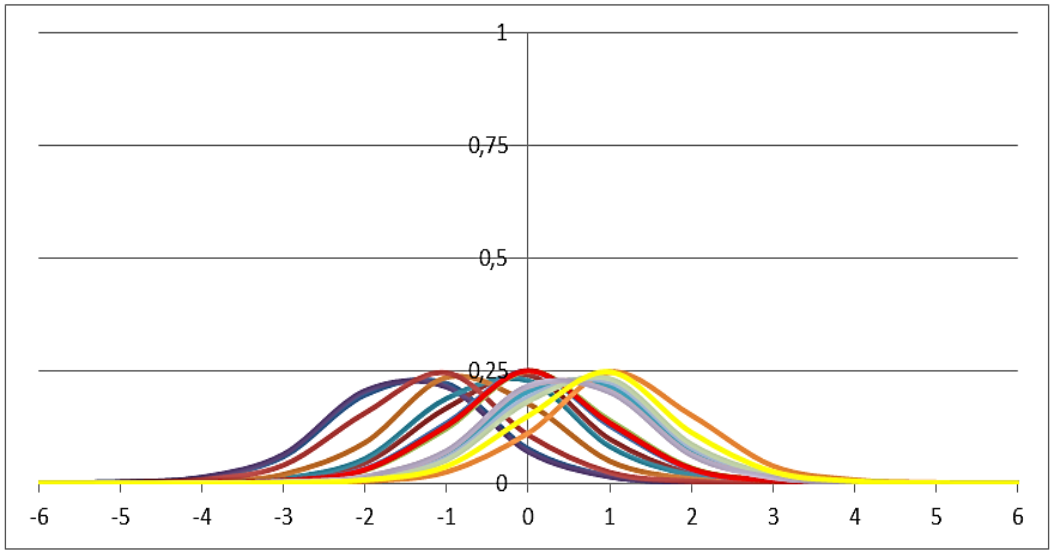

Figure 5. Item information functions for originality. In this figure, each bell-shaped line represents the information function of one item. The horizontal axis represents the latent trait continuum from -6 to 6 logits. The vertical axis shows the amount of information provided by an item.

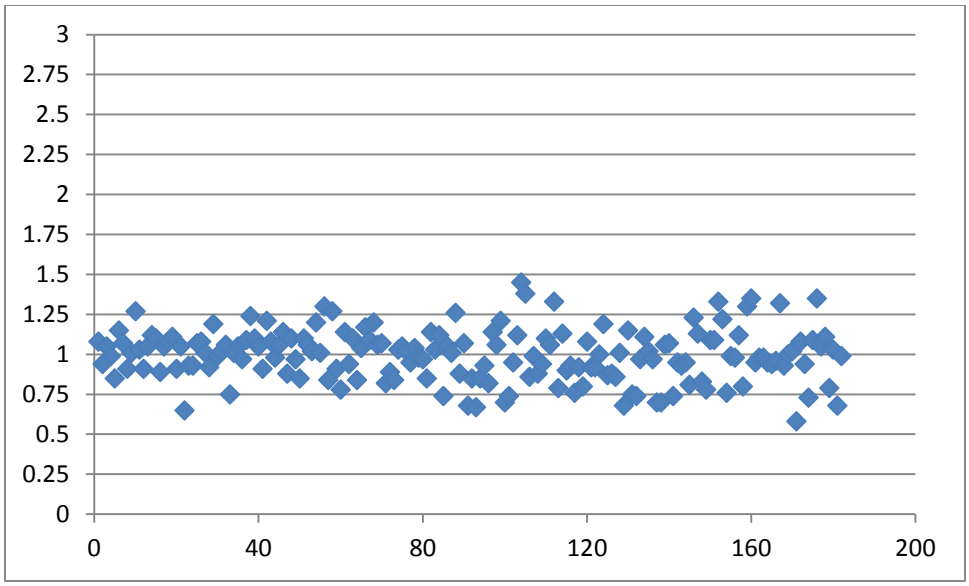

Figure 6. Person fit values for originality. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

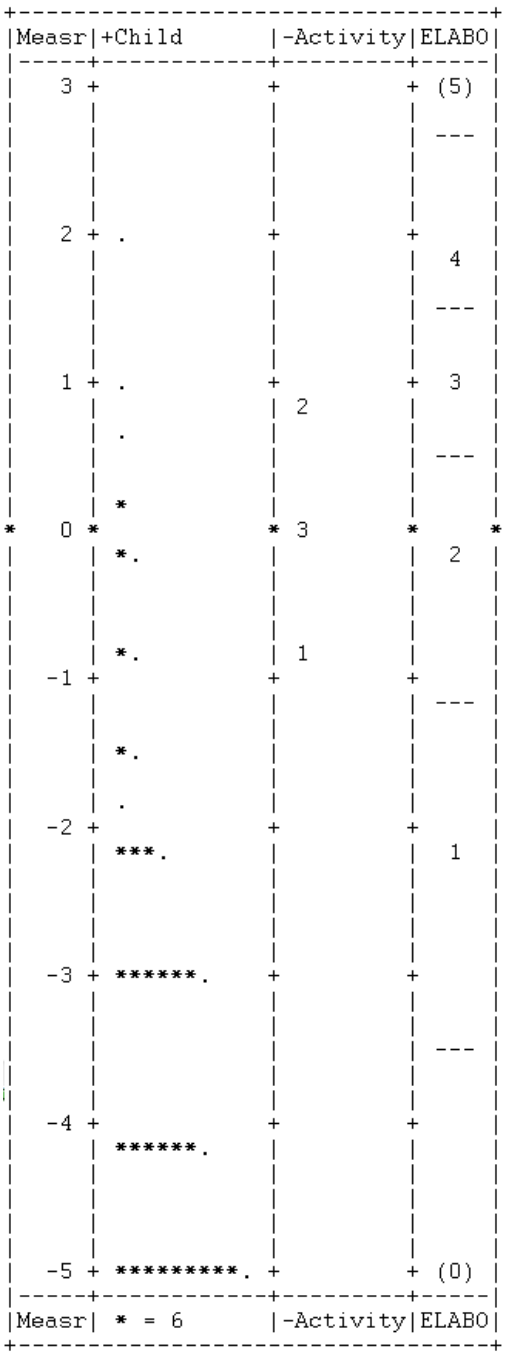

Figure 7. The variable map for elaboration. In this figure, the first column represents the latent trait continuum from -5 to 3 logits. The second column shows the locations of the students on the latent trait continuum. * represents 6 students. The third column shows the activity locations on the latent trait continuum. 1 is Activity 1. 2 is Activity 2. 3 is Activity 3. The fourth column represents the rating scale structure of a hypothetical item with the average item difficulty of 0.

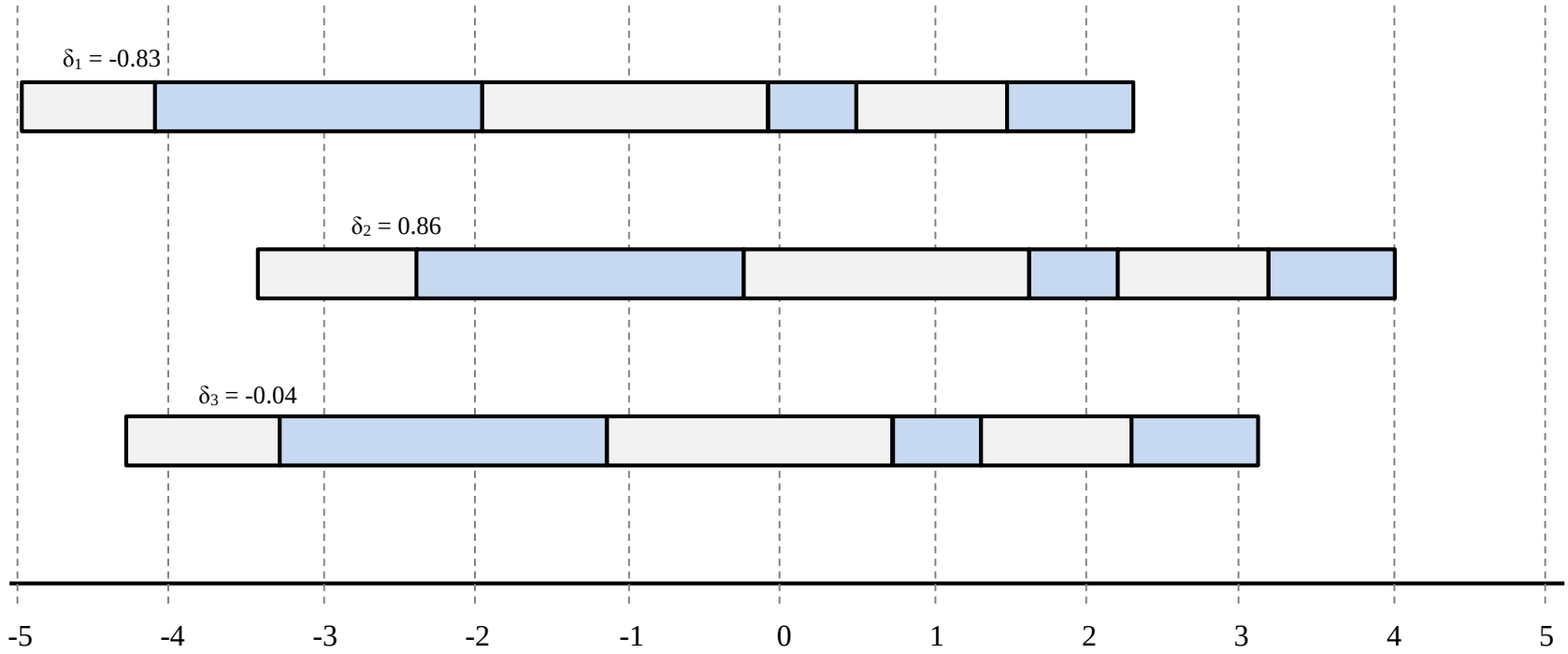

Figure 8. Rating scale structures of the activities regarding elaboration. In this figure, each bar represents one activity. Each portion in the bar represents one response option. The portion on the far left is the first response option, and the portion on the far right is the last response option. The thin line between any two portions represents the Rasch-Thurstone threshold points. The average difficulty levels of the activities are shown above the bars. The numbered line at the bottom represents the latent trait continuum from -5 to 5 logits.

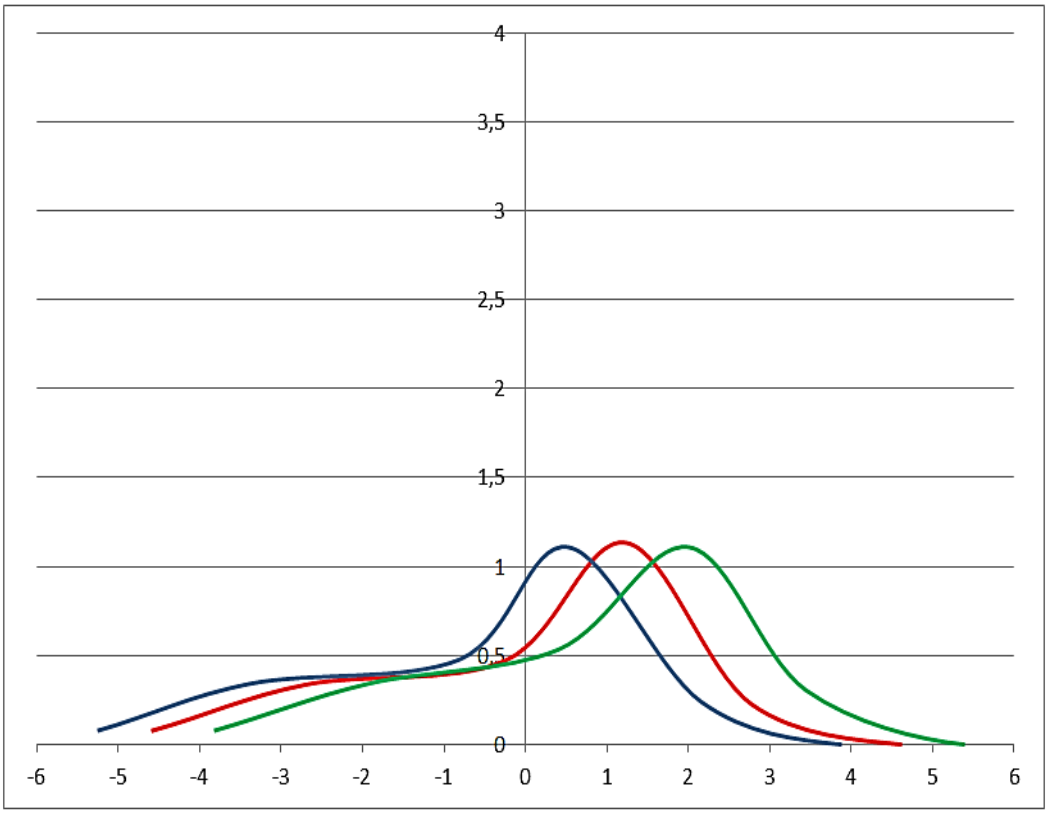

Figure 9. Information functions of the activities for elaboration. In this figure, each curved line represents the information function of one activity. The horizontal axis represents the latent trait continuum from -6 to 6 logits. The vertical axis shows the amount of information provided by an activity.

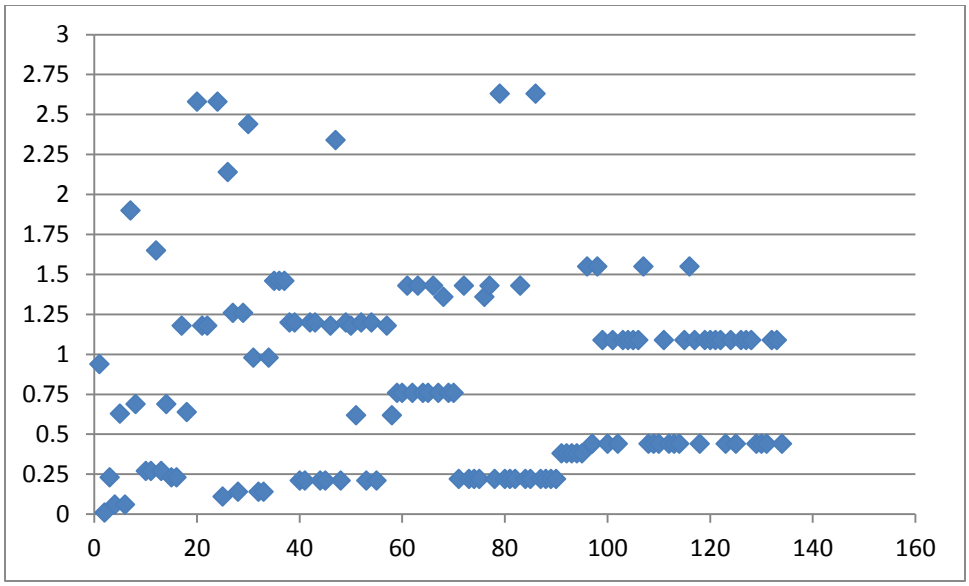

Figure 10. Person fit values for elaboration. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

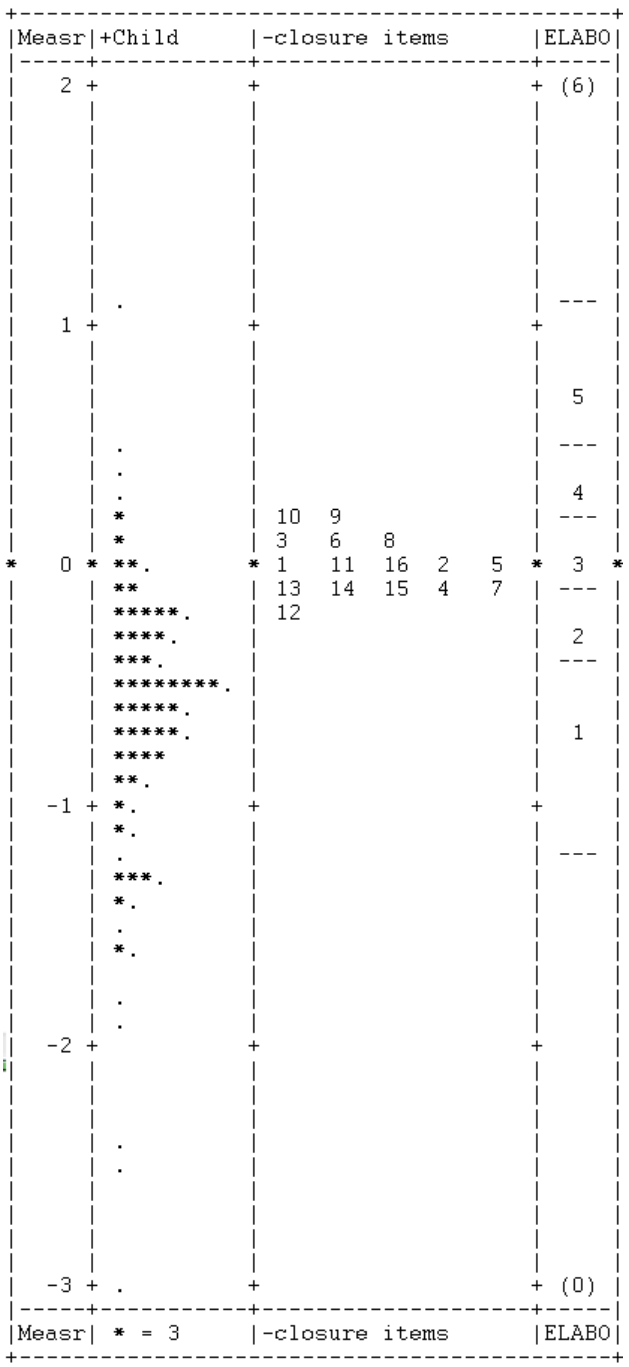

Figure 11. The variable map for elaboration-I. In this figure, the first column represents the latent trait continuum from -3 to 2 logits. The second column shows the locations of the students on the latent trait continuum. * represents 3 students. The third column shows the item locations on the latent trait continuum. 1-10 are the items in Activity 2 respectively. 11-16 are the items in Activity 3 respectively. The fourth column represents the rating scale structure of a hypothetical item with the average item difficulty of 0.

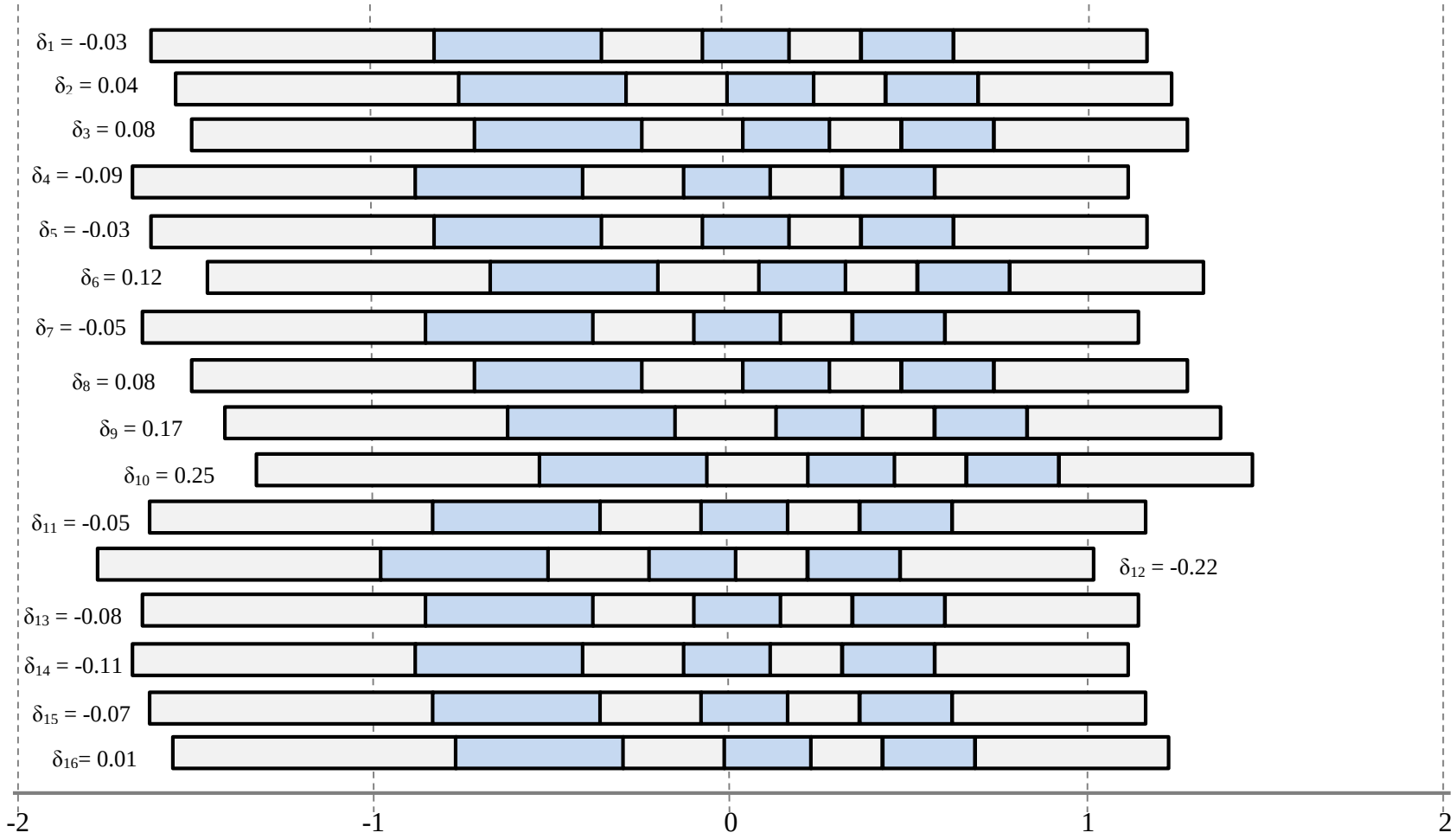

Figure 12. Rating scale structure of the items regarding elaboration-I. In this figure, each bar represents one item. Each portion in the bar represents one response option. The portion on the far left is the first response option, and the portion on the far right is the last response option. The thin line between any two portions represents the Rasch-Thurstone threshold points. The average difficulty levels of the items are shown next to the bars. The numbered line at the bottom represents the latent trait continuum from -2 to 2 logits.

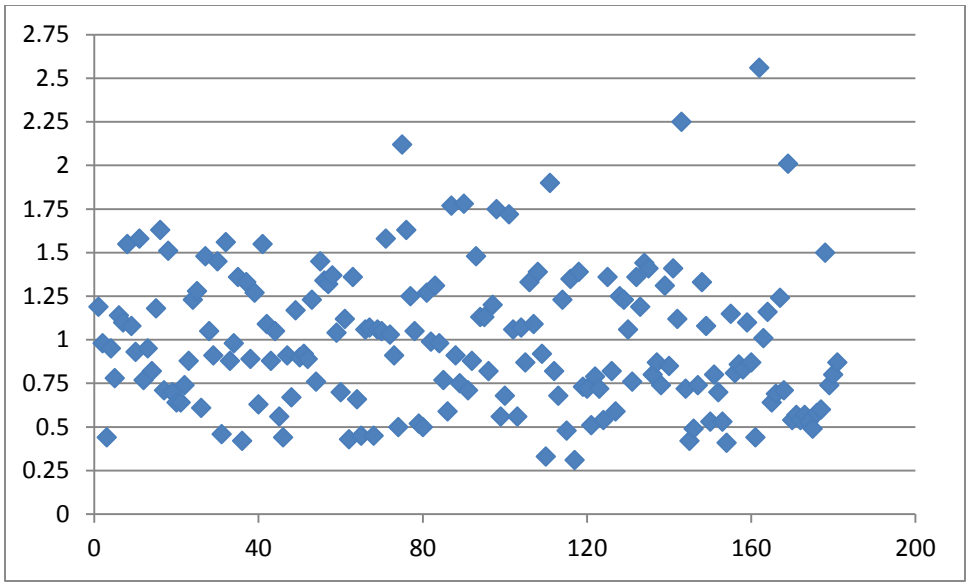

Figure 13. Person fit values for elaboration-I. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

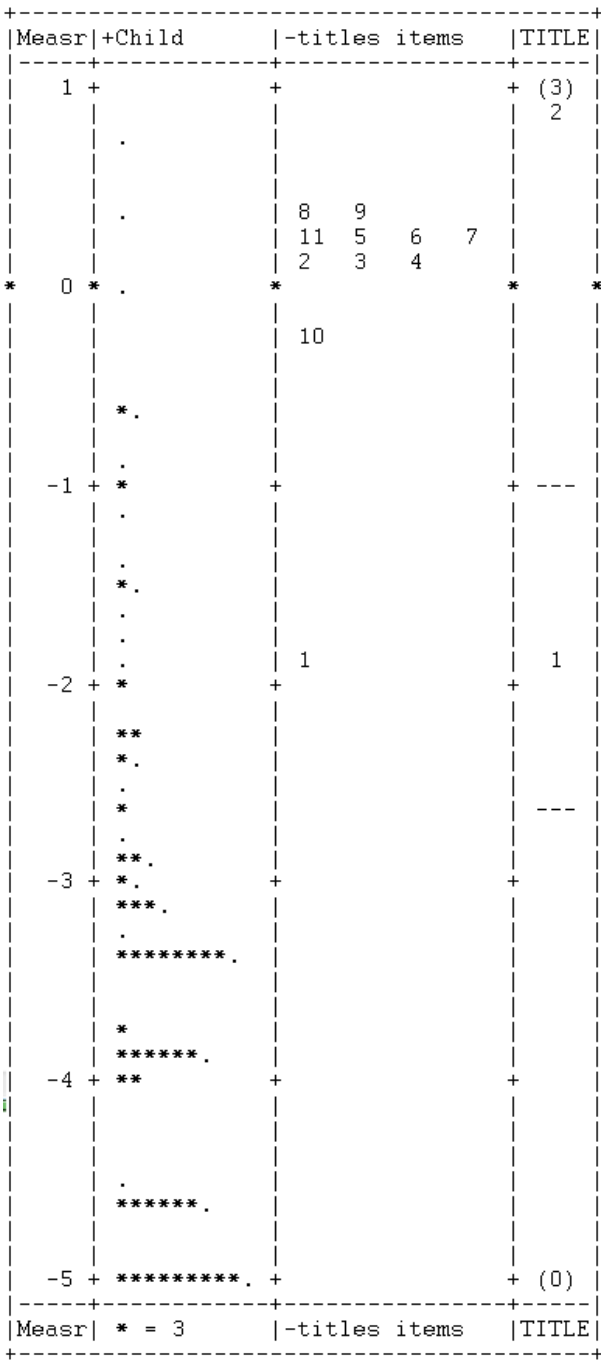

Figure 14. The variable map for abstractness of titles. In this figure, the first column represents the latent trait continuum from -5 to 1 logits. The second column shows the locations of the students on the latent trait continuum. * represents 3 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. The fourth column represents the rating scale structure of a hypothetical item with the average item difficulty of 0.

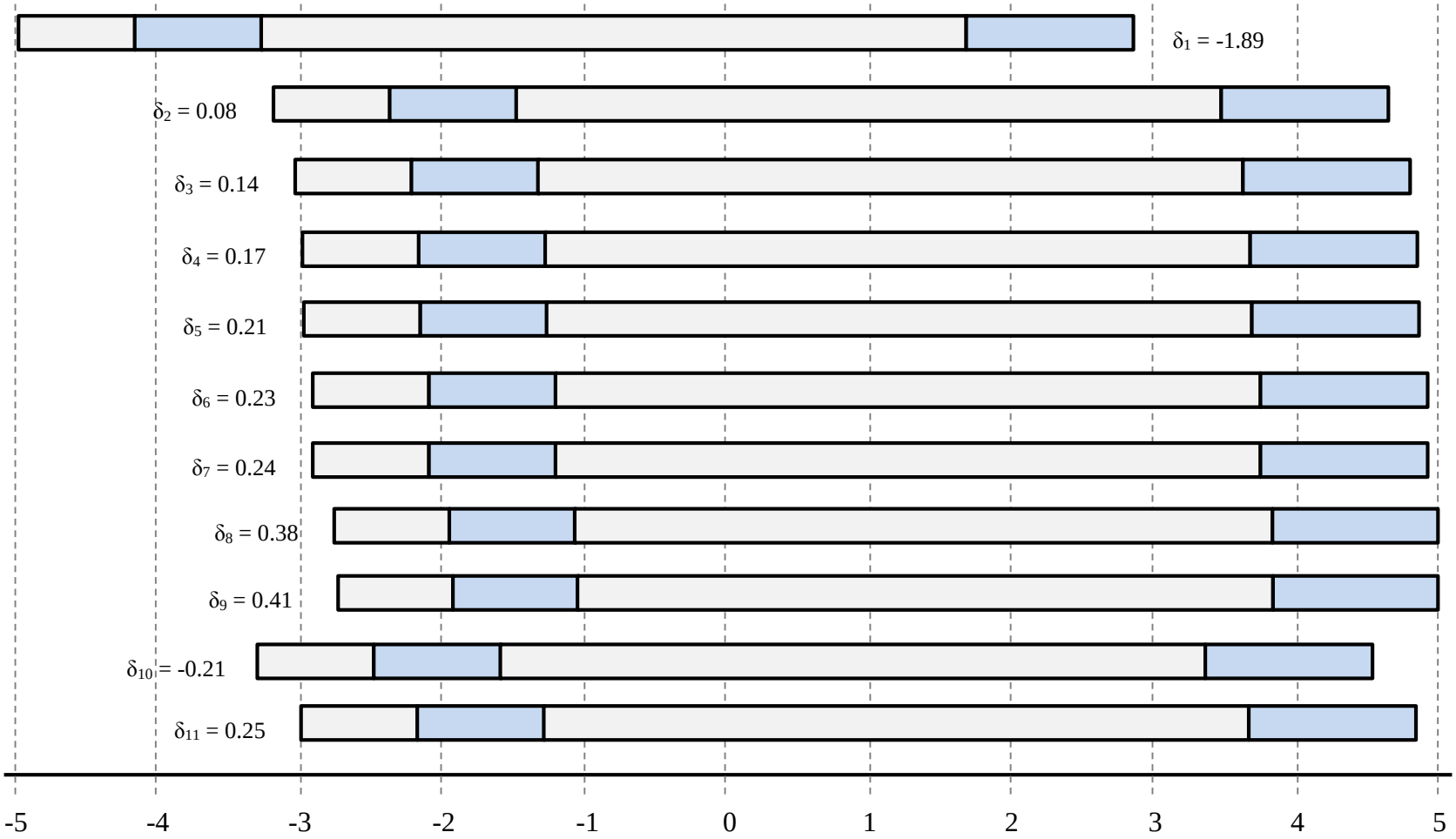

Figure 15. Rating scale structure of the items regarding abstractness of titles. In this figure, each bar represents one item. Each portion in the bar represents one response option. The portion on the far left is the first response option, and the portion on the far right is the last response option. The thin line between any two portions represents the Rasch-Thurstone threshold points. The average difficulty levels of the items are shown next to the bars. Numbered line at the bottom represents the latent trait continuum from -5 to 5 logits.

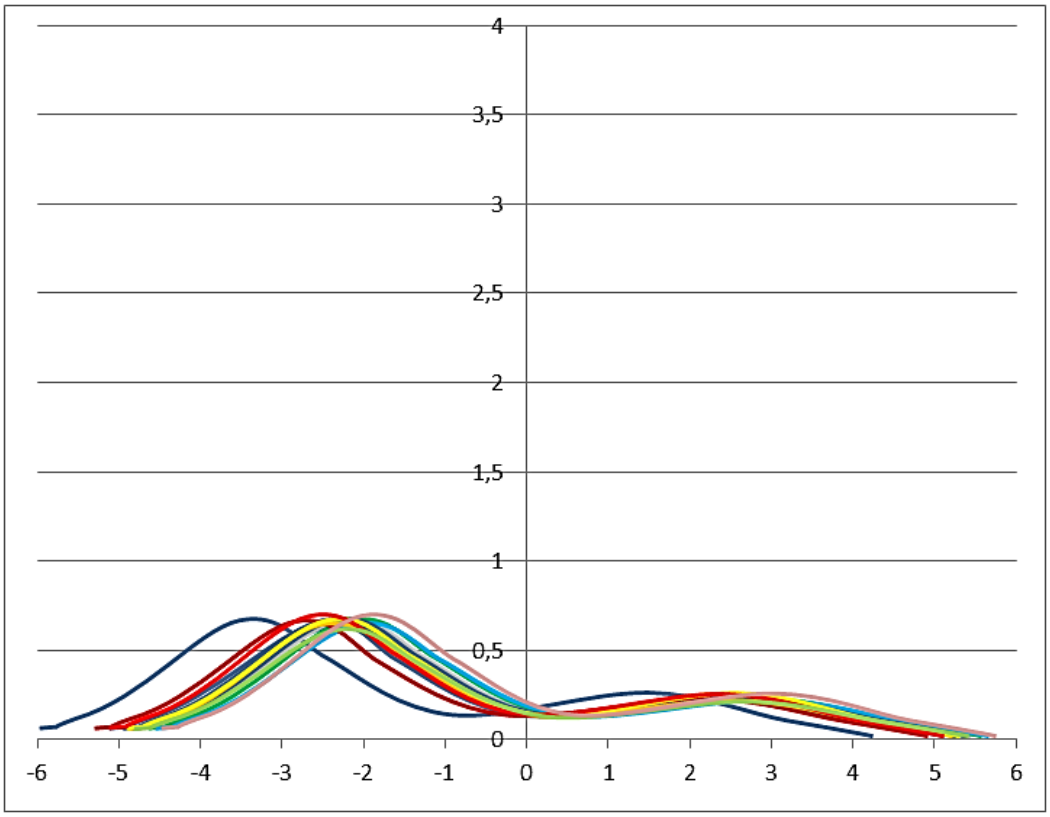

Figure 16. Item information functions for abstractness of titles. In this figure, each curved line represents the information function of one item. The horizontal axis represents the latent trait continuum from -6 to 6 logits. The vertical axis shows the amount of information provided by an item.

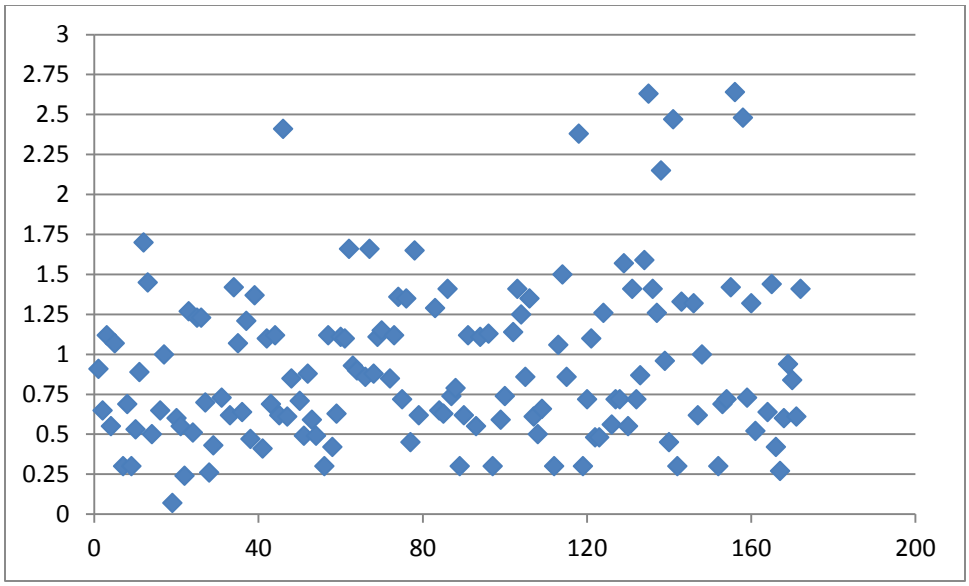

Figure 17. Person fit values for abstractness of titles. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

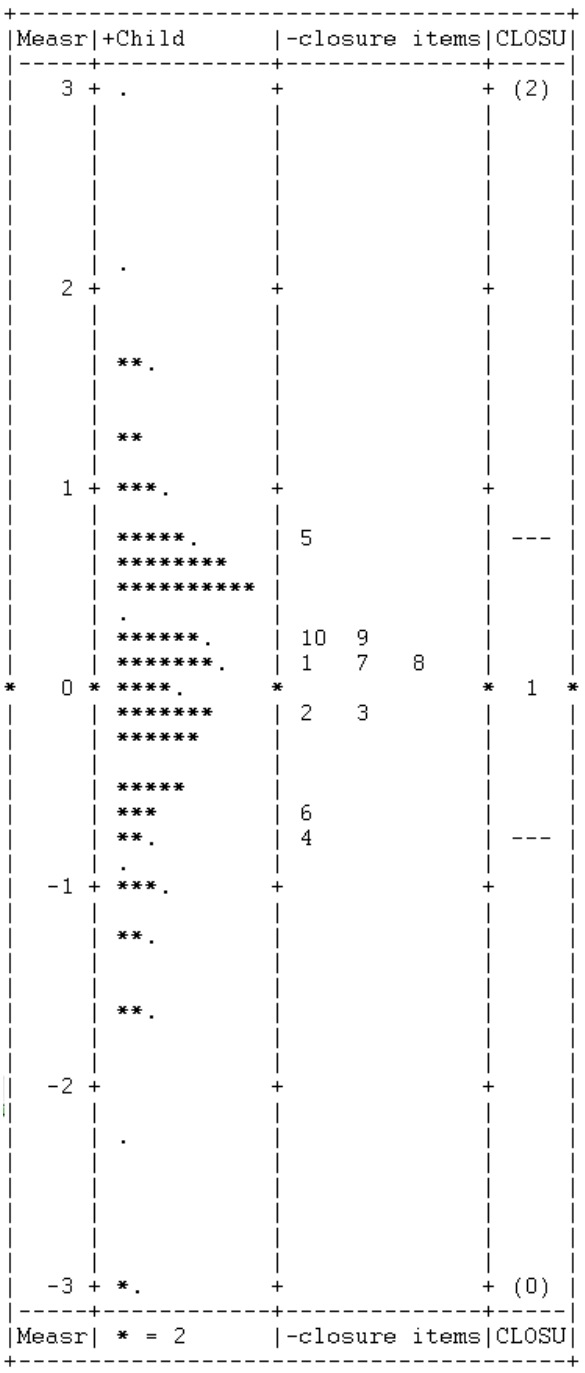

Figure 18. The variable map for resistance to premature closure. In this figure, the first column represents the latent trait continuum from -3 to 3 logits. The second column shows the locations of the students on the latent trait continuum. * represents 2 students. The third column shows the item locations on the latent trait continuum. 1-10 are the items in Activity 2 respectively. The fourth column represents the rating scale structure of a hypothetical item with the average item difficulty of 0.

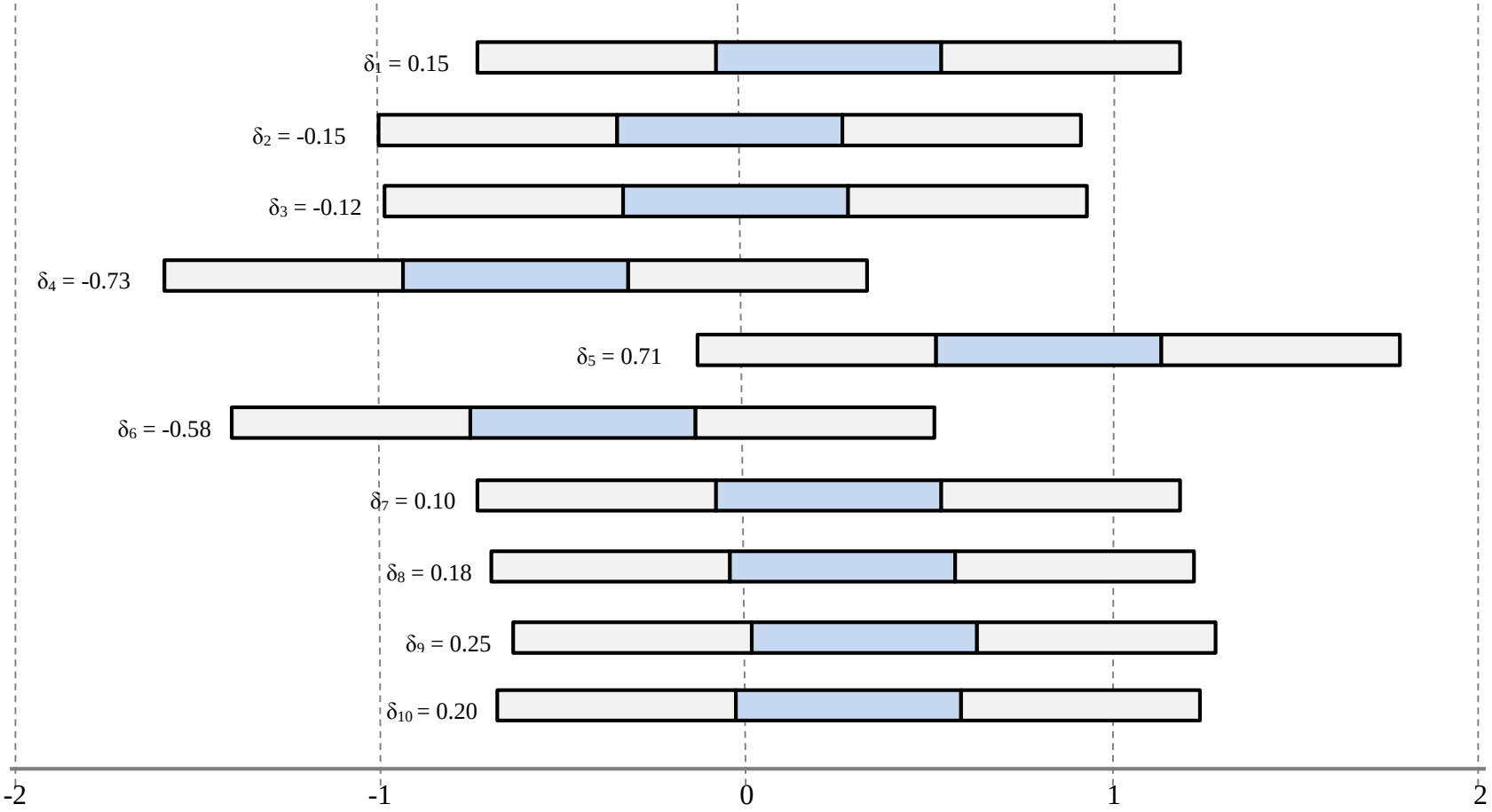

Figure 19. Rating scale structure of the items regarding resistance to premature closure. In this figure, each bar represents one item. Each portion in the bar represents one response option. The portion on the far left is the first response option, and the portion on the far right is the last response option. The thin line between any two portions represents the Rasch-Thurstone threshold points. The average difficulty levels of the items are shown next to the bars. The numbered line at the bottom represents the latent trait continuum from -2 to 2 logits.

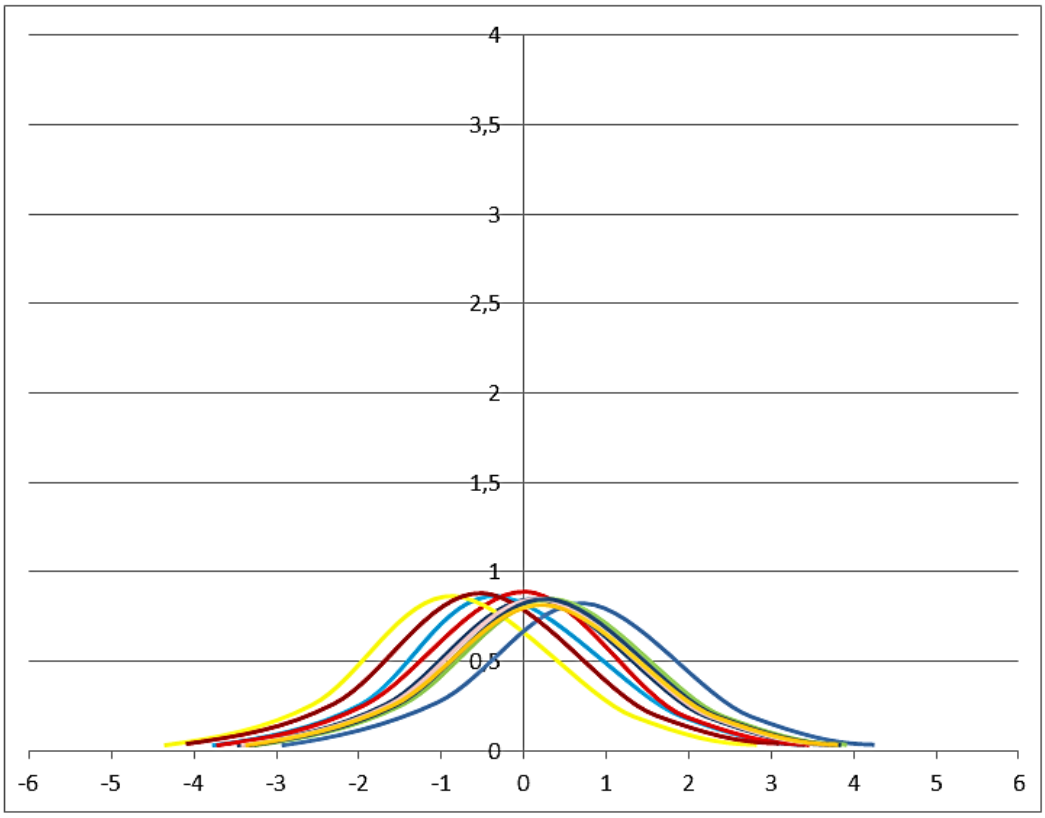

Figure 20. Item information functions for resistance to premature closure. In this figure, each bell-shaped line represents the information function of one item. The horizontal axis represents the latent trait continuum from -6 to 6 logits. The vertical axis shows the amount of information provided by an item.

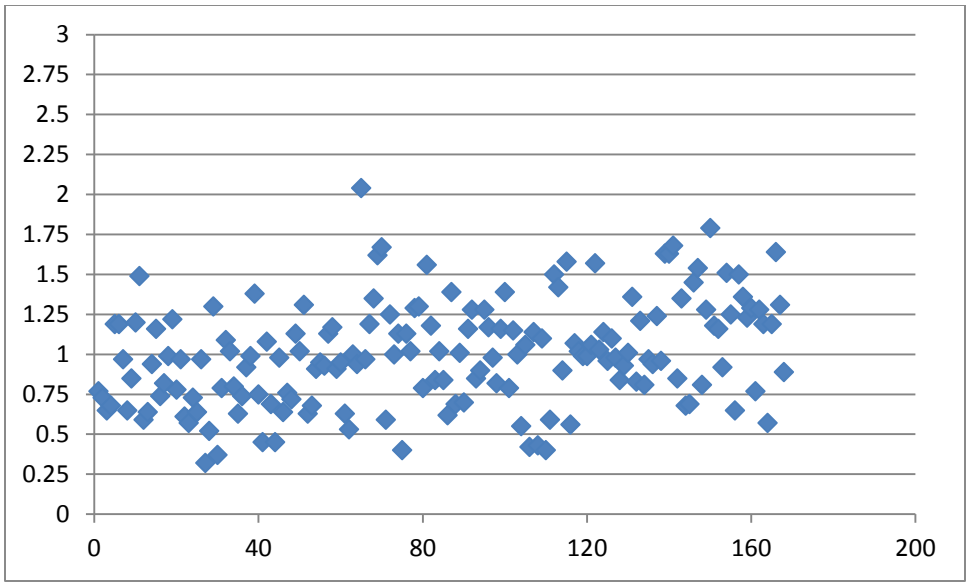

Figure 21. Person fit values for resistance to premature closure. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

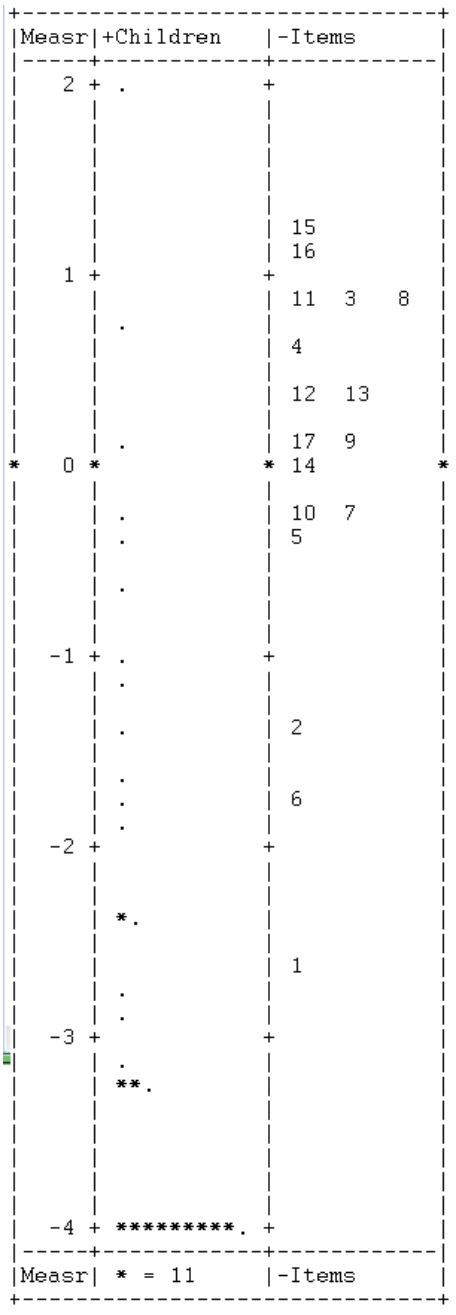

Figure 22. The variable map for emotional expressiveness. In this figure, the first column represents the latent trait continuum from -4 to 2 logits. The second column shows the locations of the students on the latent trait continuum. * represents 11 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

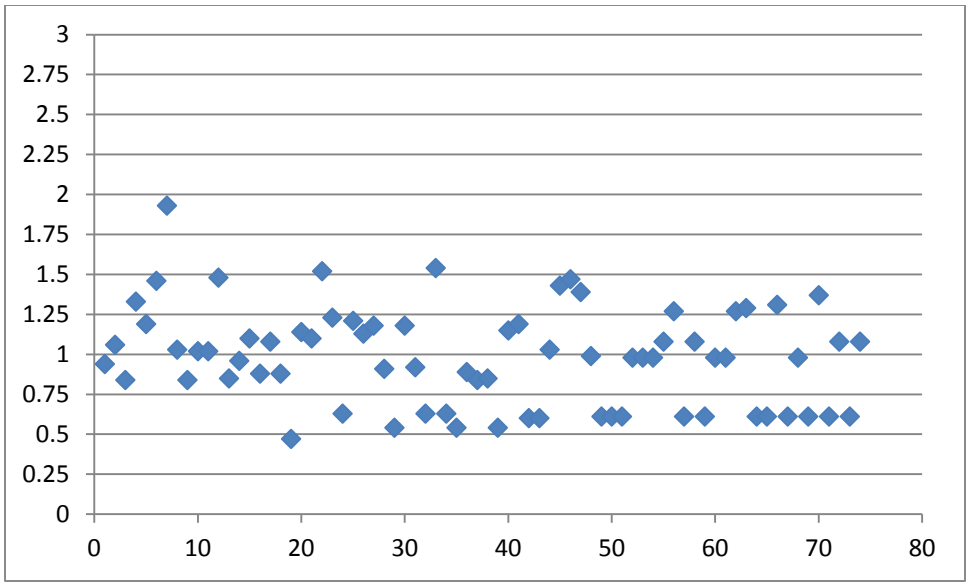

Figure 23. Person fit values for emotional expressiveness. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

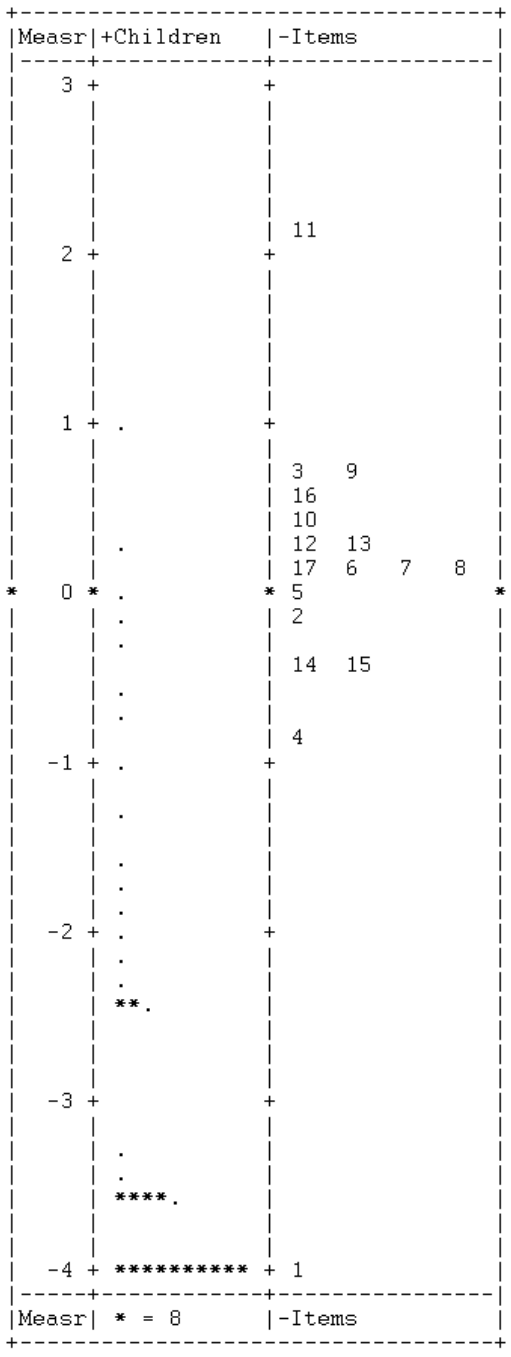

Figure 24. The variable map for storytelling articulateness. In this figure, the first column represents the latent trait continuum from -4 to 3 logits. The second column shows the locations of the students on the latent trait continuum. * represents 8 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

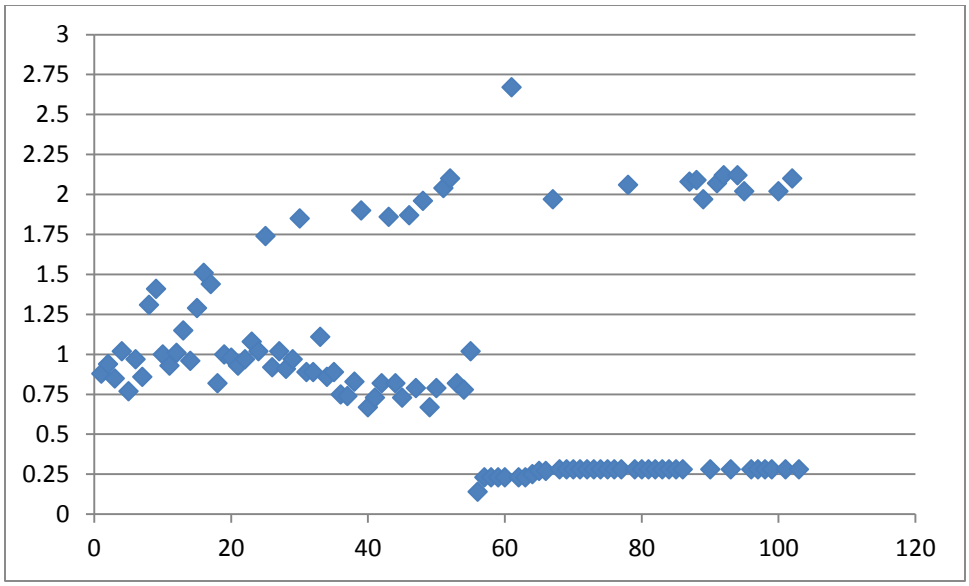

Figure 25. Person fit values for storytelling articulateness. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

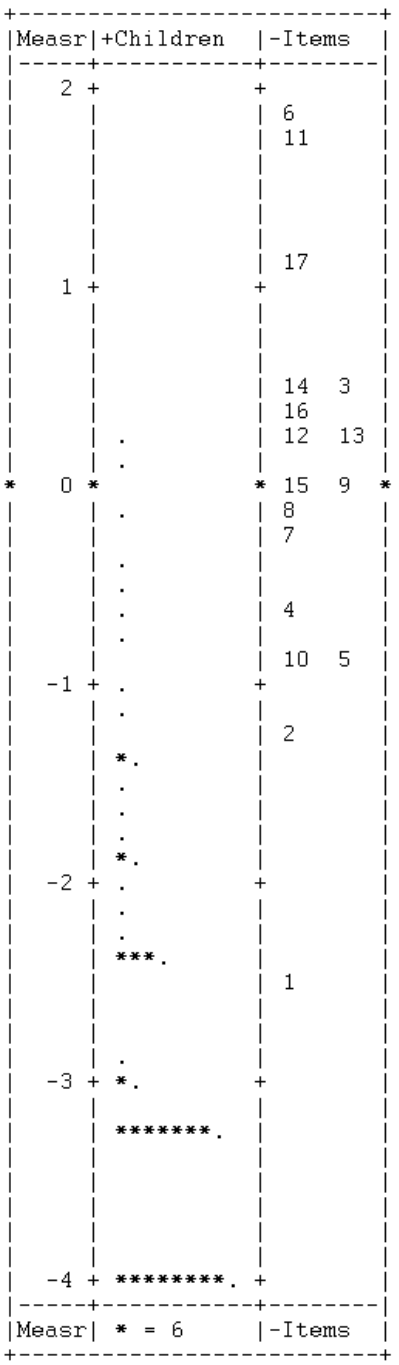

Figure 26. The variable map for movement or action. In this figure, the first column represents the latent trait continuum from -4 to 2 logits. The second column shows the locations of the students on the latent trait continuum. * represents 6 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

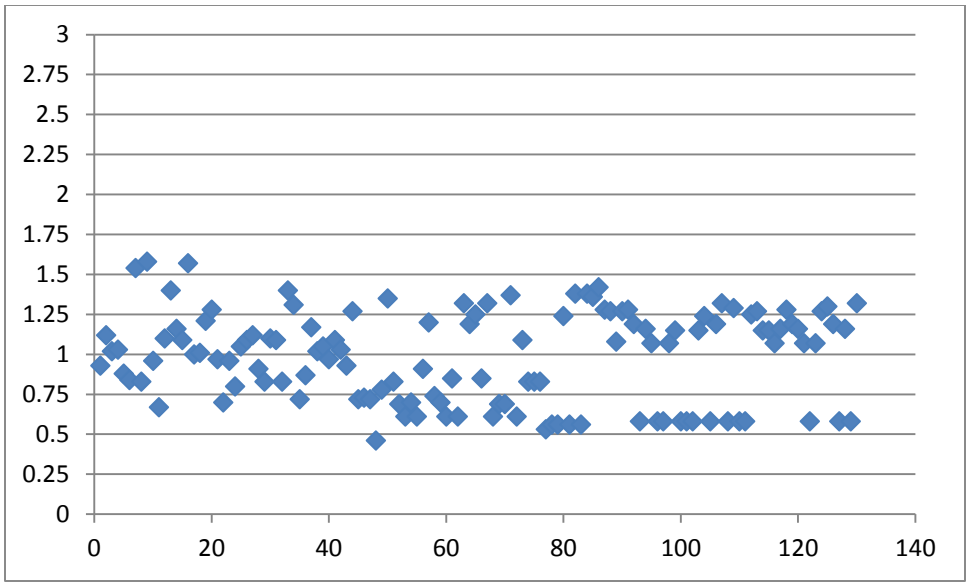

Figure 27. Person fit values for movement or action. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

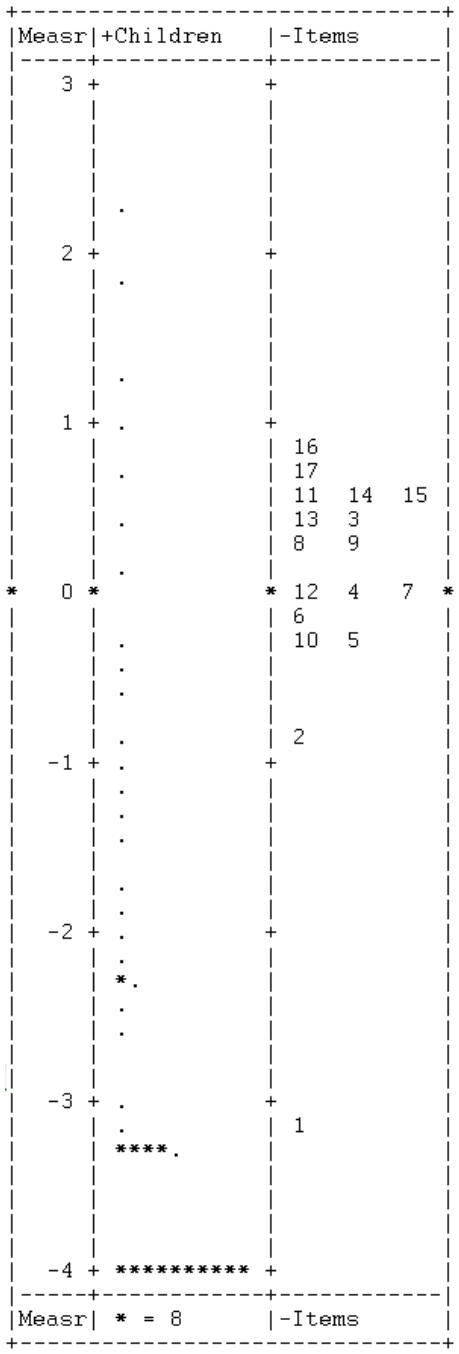

Figure 28. The variable map for expressiveness of titles. In this figure, the first column represents the latent trait continuum from -4 to 3 logits. The second column shows the locations of the students on the latent trait continuum. * represents 8 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

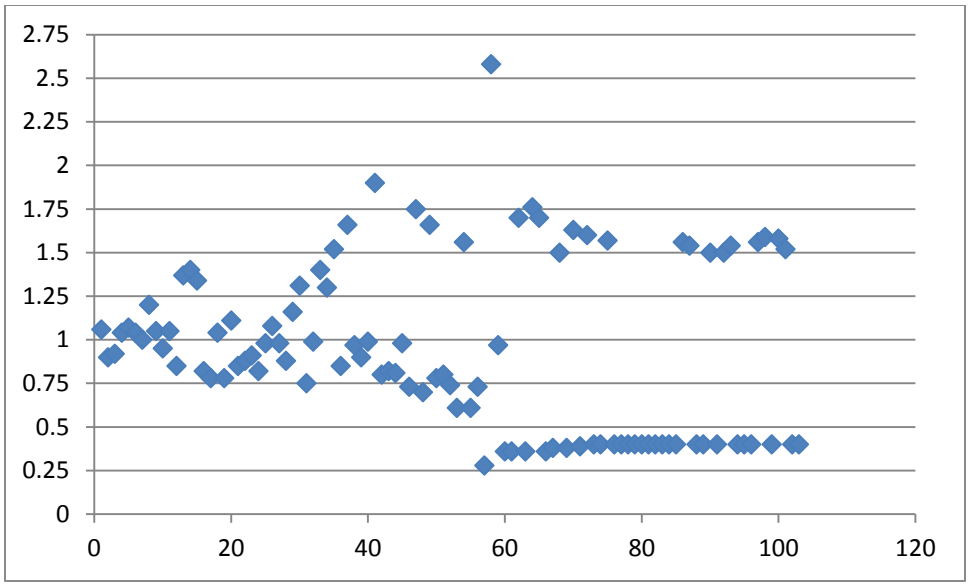

Figure 29. Person fit values for expressiveness of titles. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

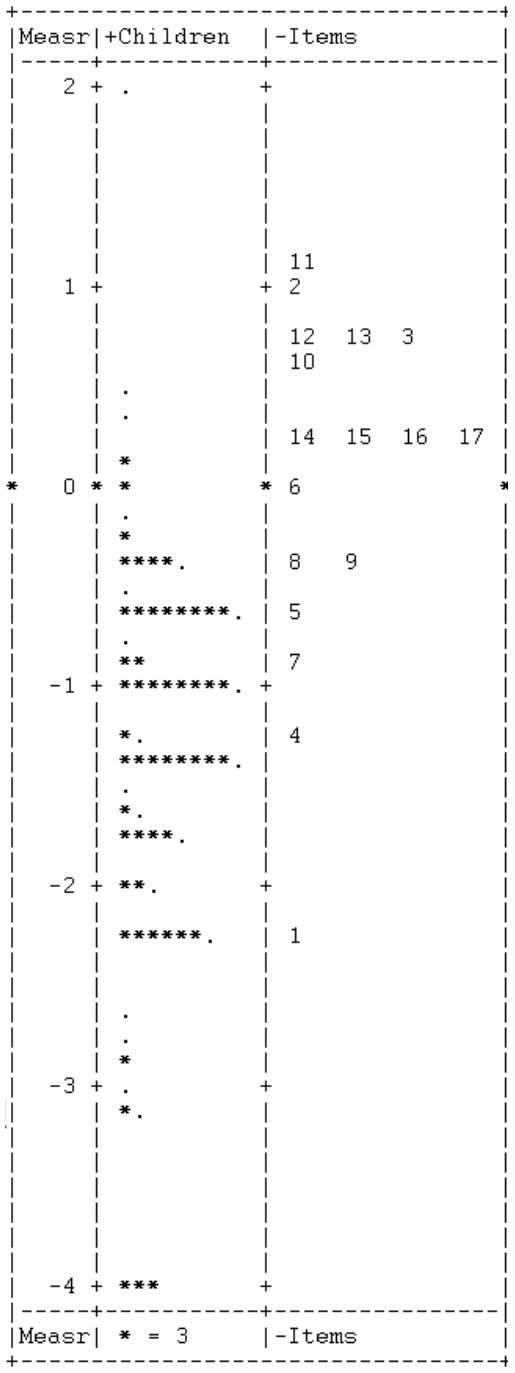

Figure 30. The variable map for unusual visualization. In this figure, the first column represents the latent trait continuum from -4 to 2 logits. The second column shows the locations of the students on the latent trait continuum. * represents 3 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

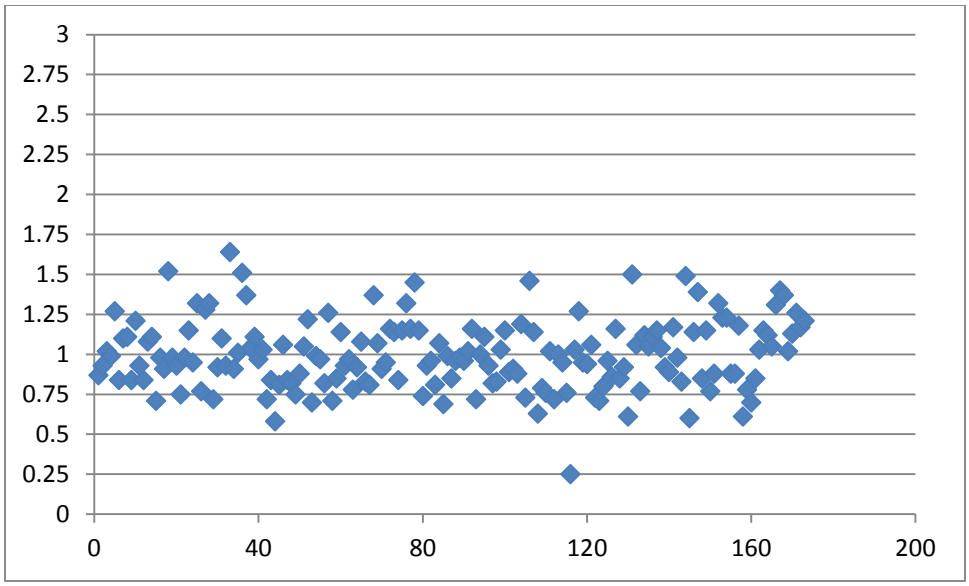

Figure 31. Person fit values for unusual visualization. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

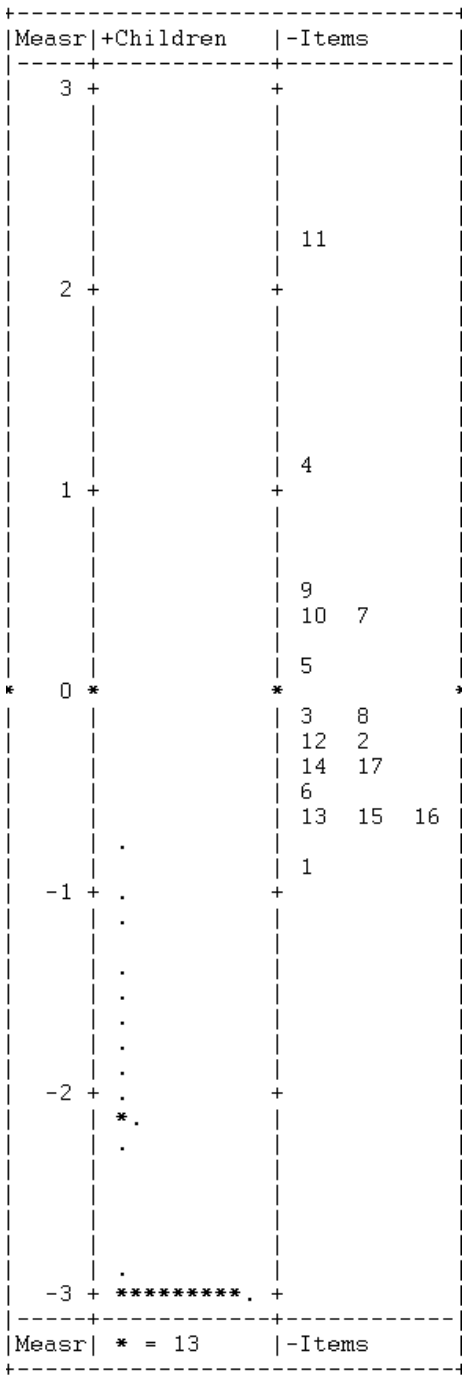

Figure 32. The variable map for internal visualization. In this figure, the first column represents the latent trait continuum from -3 to 3 logits. The second column shows the locations of the students on the latent trait continuum. * represents 13 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

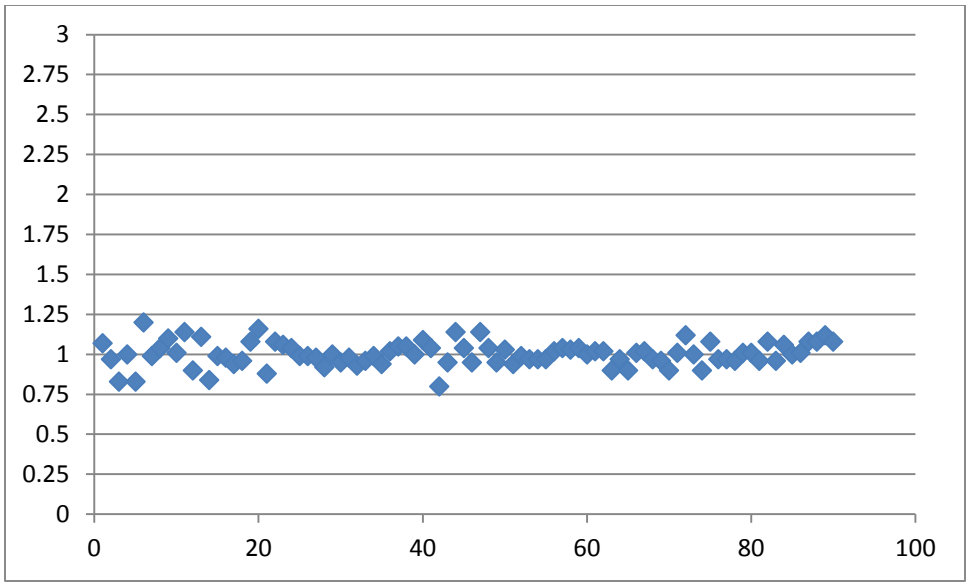

Figure 33. Person fit values for internal visualization. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

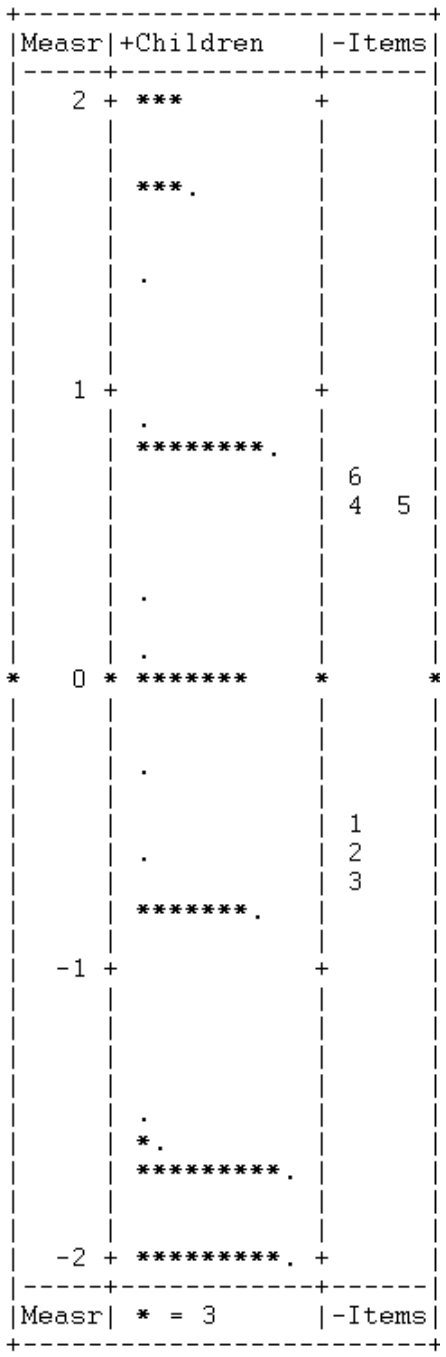

Figure 34. The variable map for extending or breaking boundaries. In this figure, the first column represents the latent trait continuum from -2 to 2 logits. The second column shows the locations of the students on the latent trait continuum. * represents 3 students. The third column shows the item locations on the latent trait continuum. 1-6 are the first six items in Activity 3 respectively.

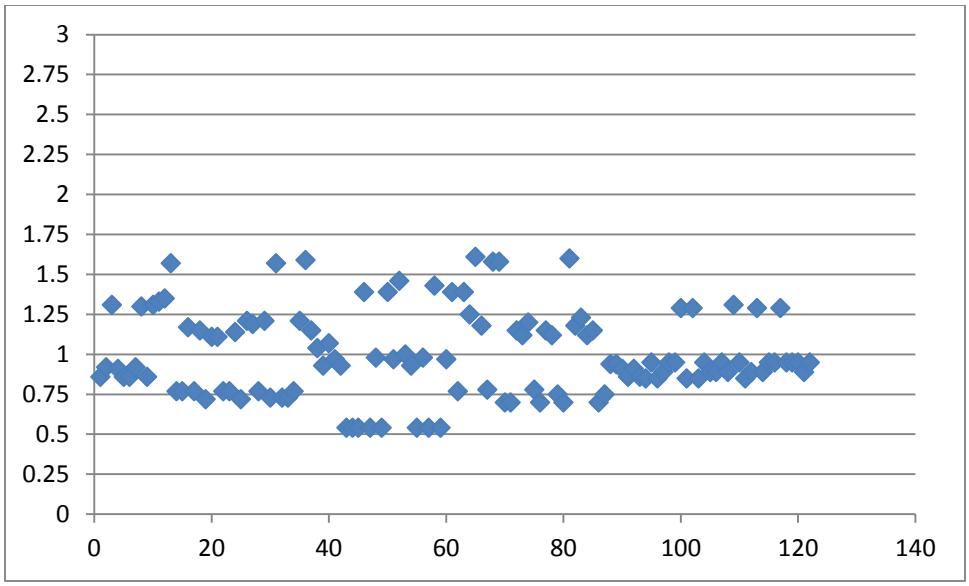

Figure 35. Person fit values for extending or breaking boundaries. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

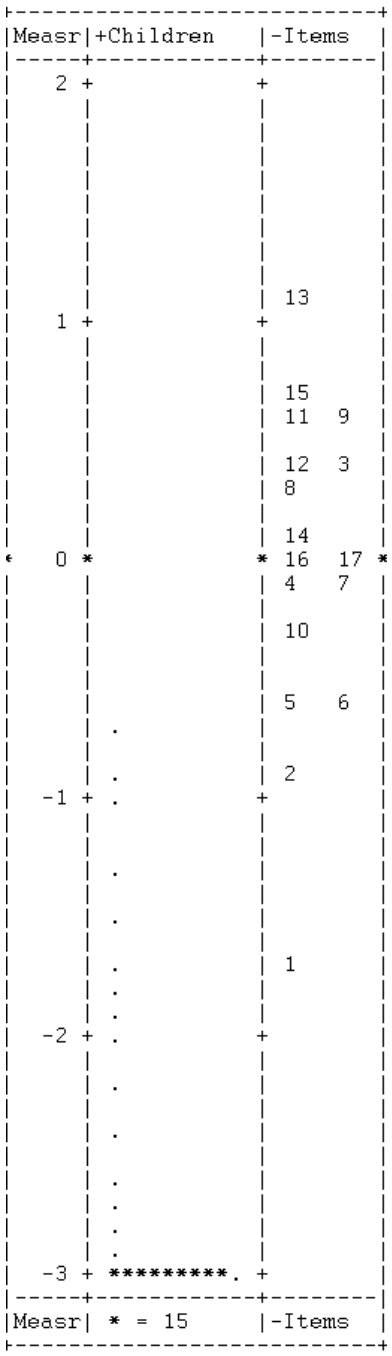

Figure 36. The variable map for humor. In this figure, the first column represents the latent trait continuum from -3 to 2 logits. The second column shows the locations of the students on the latent trait continuum. * represents 15 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

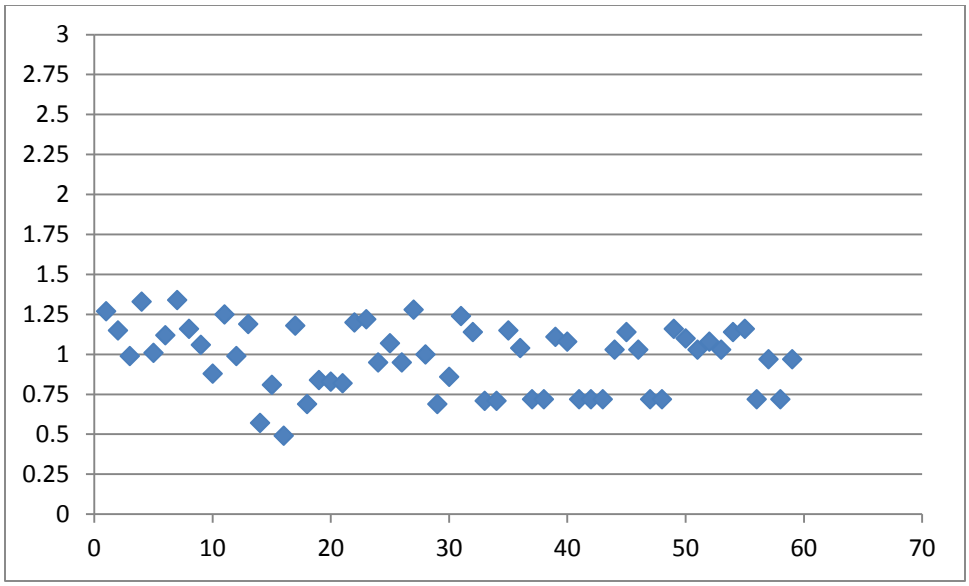

Figure 37. Person fit values for humor. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

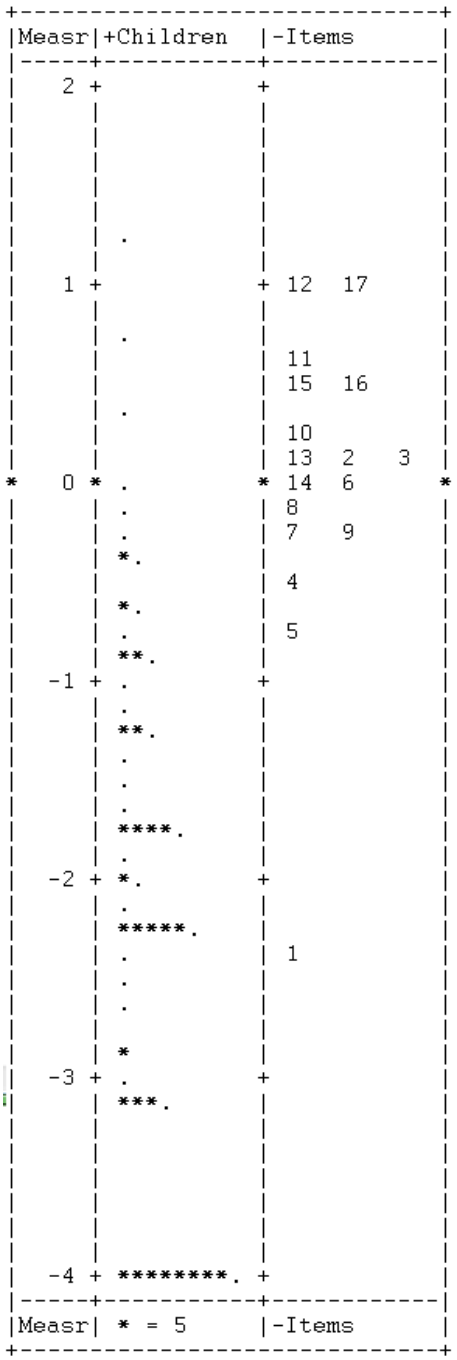

Figure 38. The variable map for richness of imagery. In this figure, the first column represents the latent trait continuum from -4 to 2 logits. The second column shows the locations of the students on the latent trait continuum. * represents 5 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

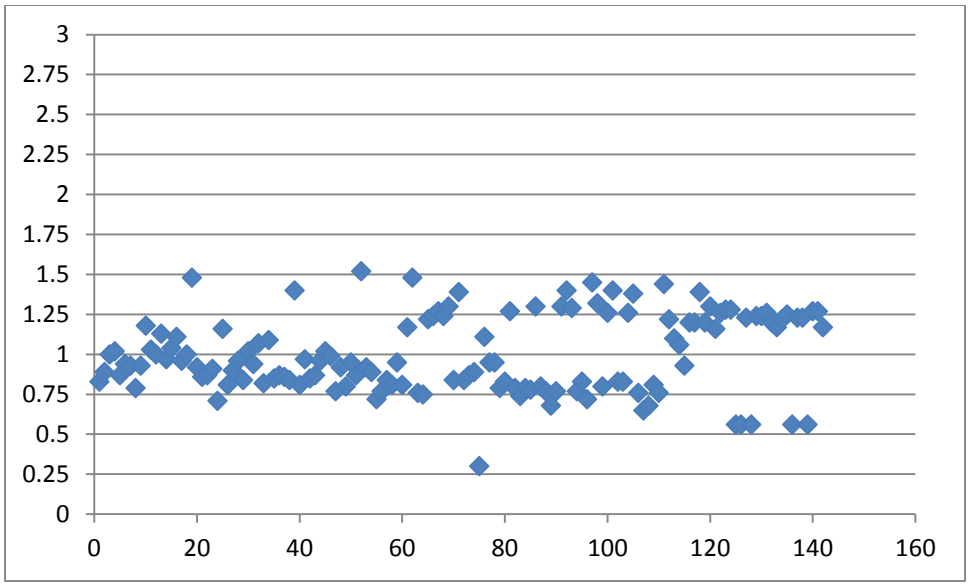

Figure 39. Person fit values for richness of imagery. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

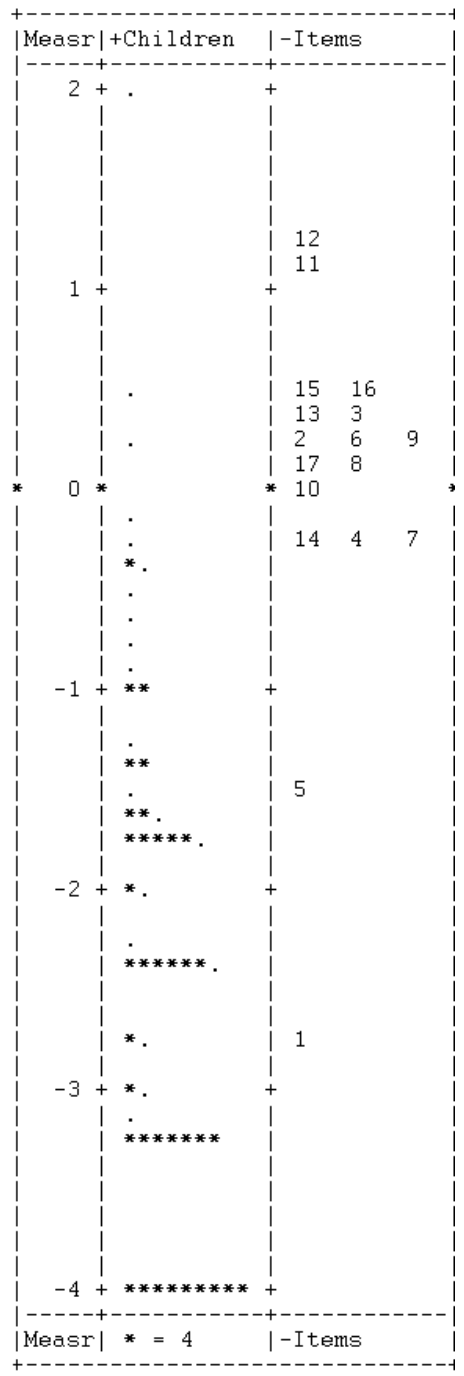

Figure 40. The variable map for colorfulness of imagery. In this figure, the first column represents the latent trait continuum from -4 to 2 logits. The second column shows the locations of the students on the latent trait continuum. * represents 4 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

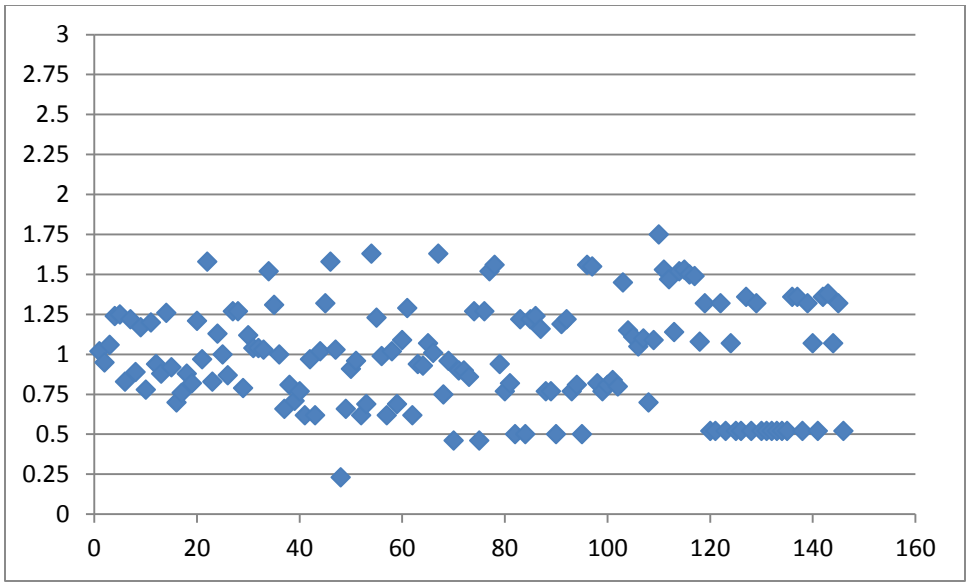

Figure 41. Person fit values for colorfulness of imagery. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

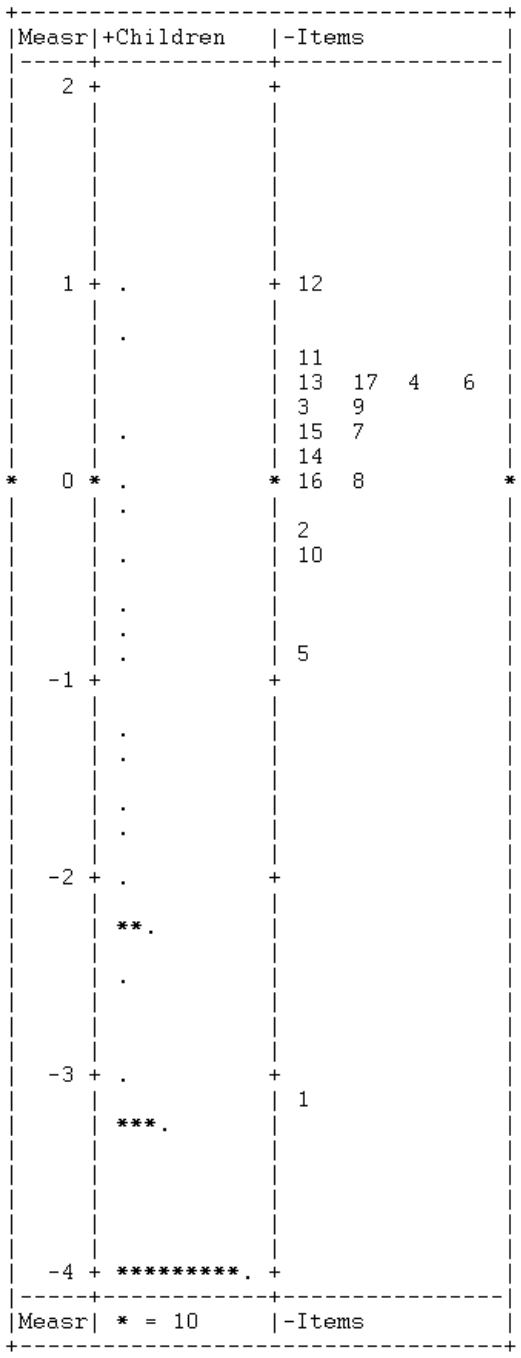

Figure 42. The variable map for fantasy. In this figure, the first column represents the latent trait continuum from -4 to 2 logits. The second column shows the locations of the students on the latent trait continuum. * represents 10 students. The third column shows the item locations on the latent trait continuum. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the items in Activity 3 respectively.

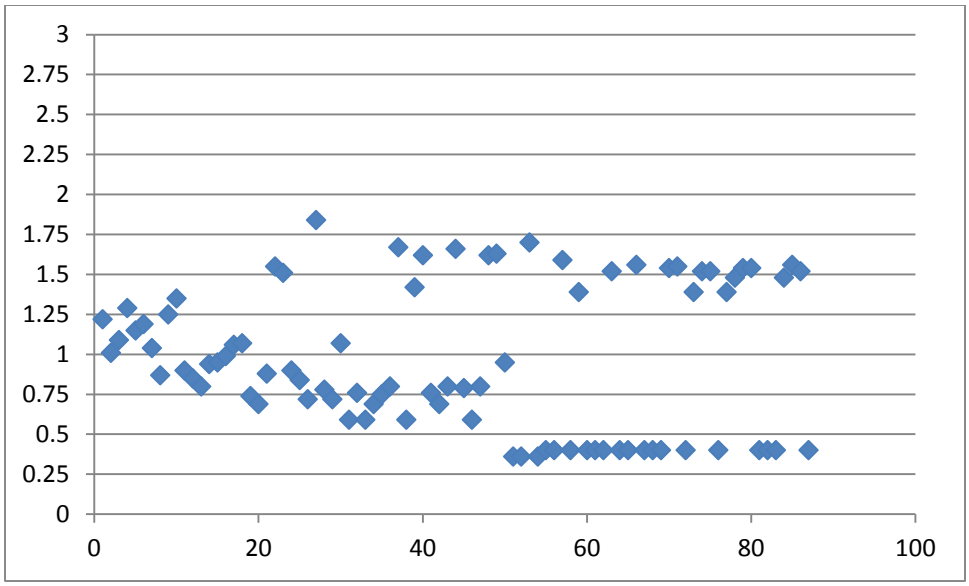

Figure 43. Person fit values for fantasy. This figure shows the infit mean square values for the students. Each spot represents one student. The horizontal axis shows the number of students. The vertical axis shows the values for person fit.

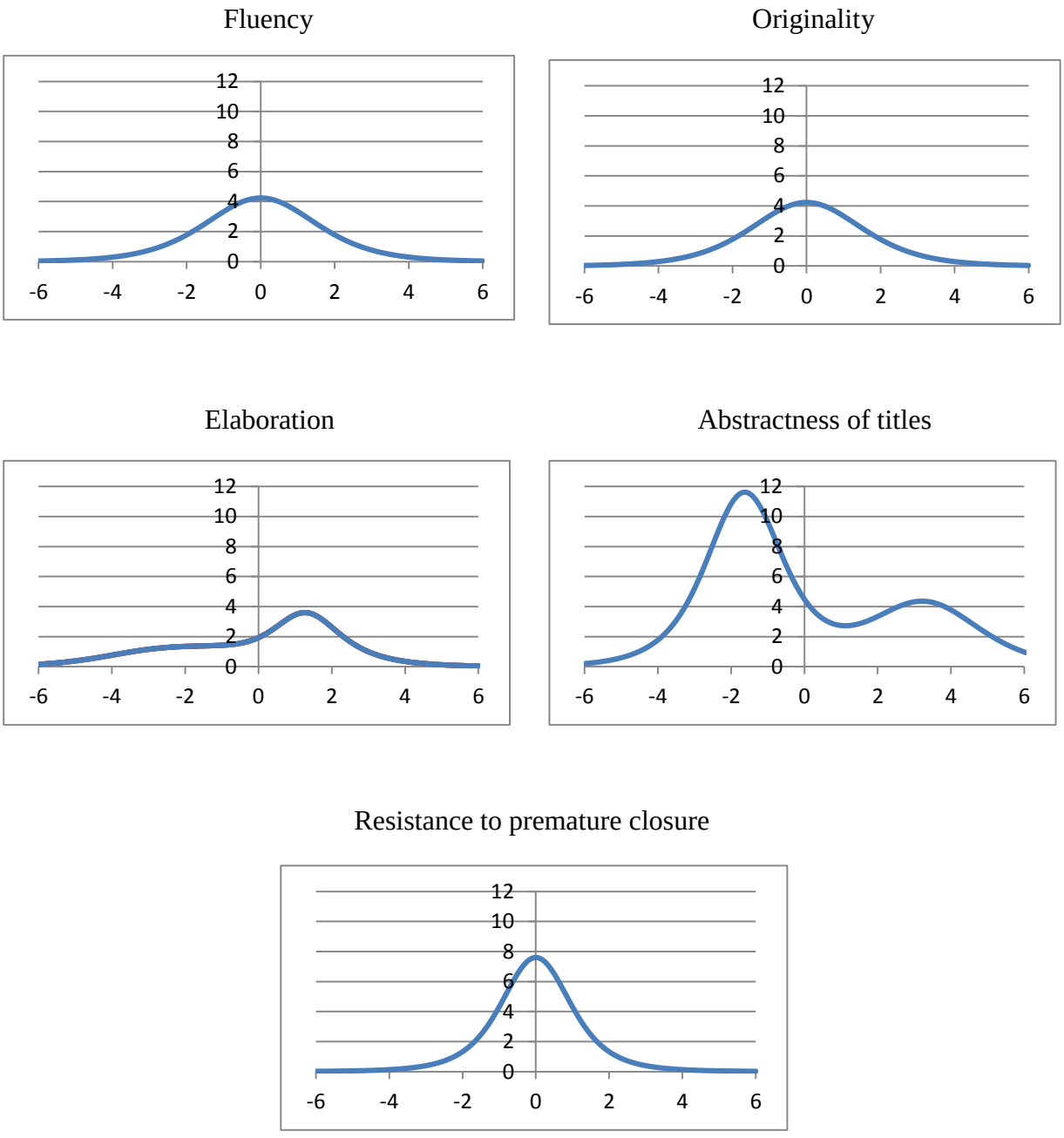

Figure 44. Test information functions for the norm-referenced variables. In this figure, each curved line represents the test information function of one variable. The horizontal axis represents the latent trait continuum from -6 to 6 logits. The vertical axis shows the total amount of information provided by the items.

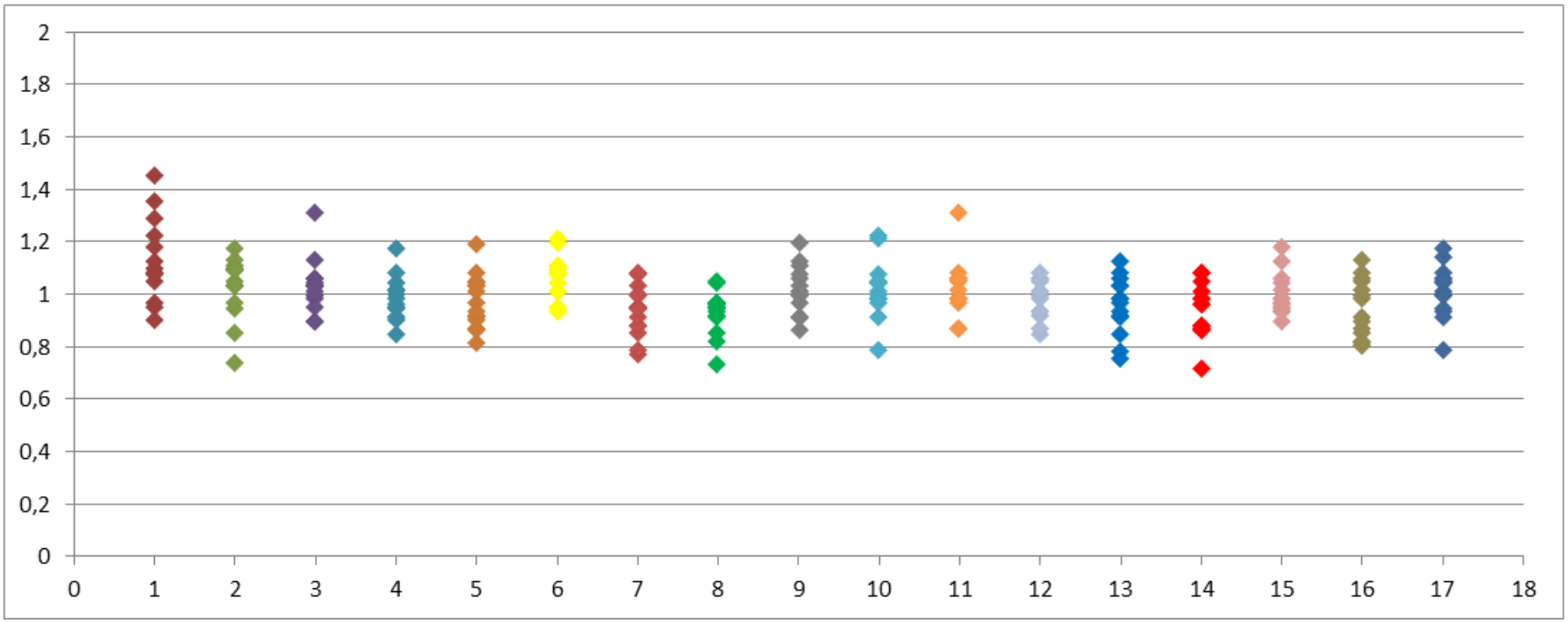

Figure 45. Infit mean square values for the items regarding each variable. The numbered line at the bottom shows the item numbers from 1 to 17. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the first six items in Activity 3 respectively. The vertical axis shows the values for item fit. Each colored spot represents one variable measured by the item.

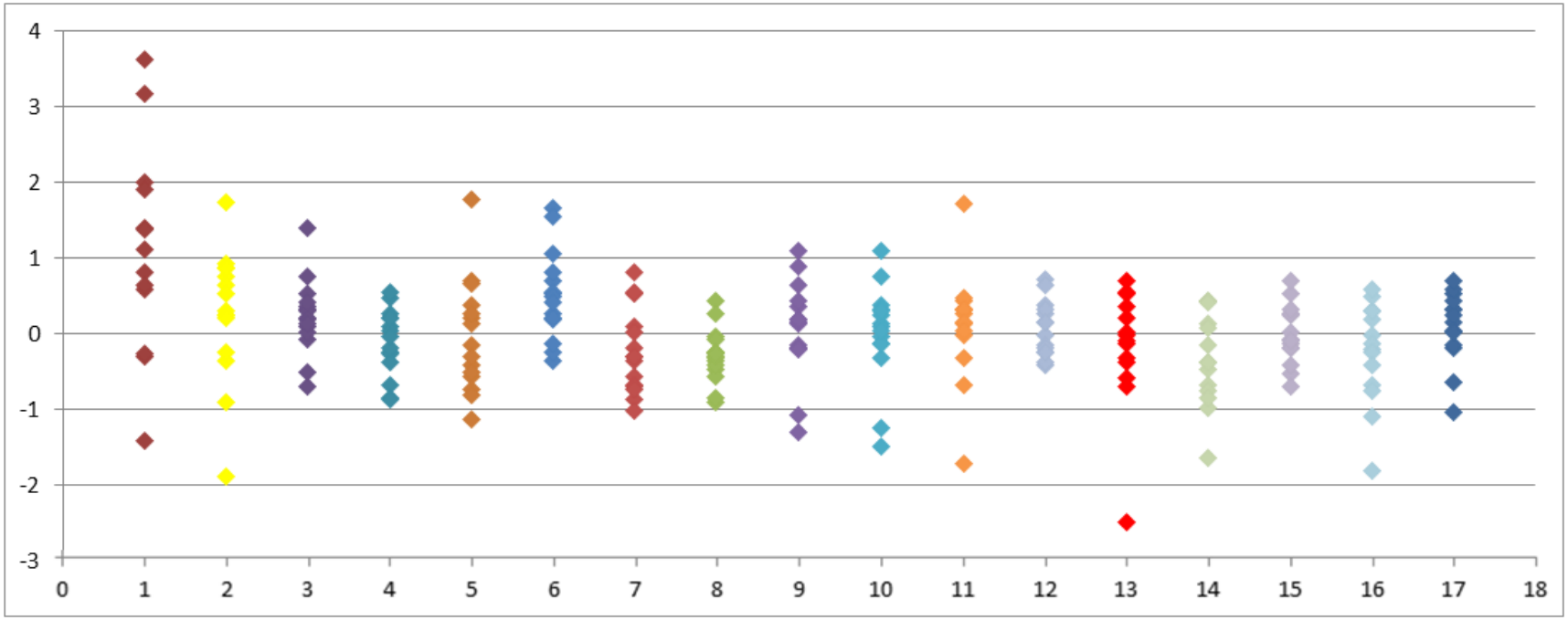

Figure 46. Standardized infit values for the items regarding each variable. The numbered line at the bottom shows the item numbers from 1 to 17. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the first six items in Activity 3 respectively. The vertical axis shows the values for item fit. Each colored spot represents one variable measured by the item.

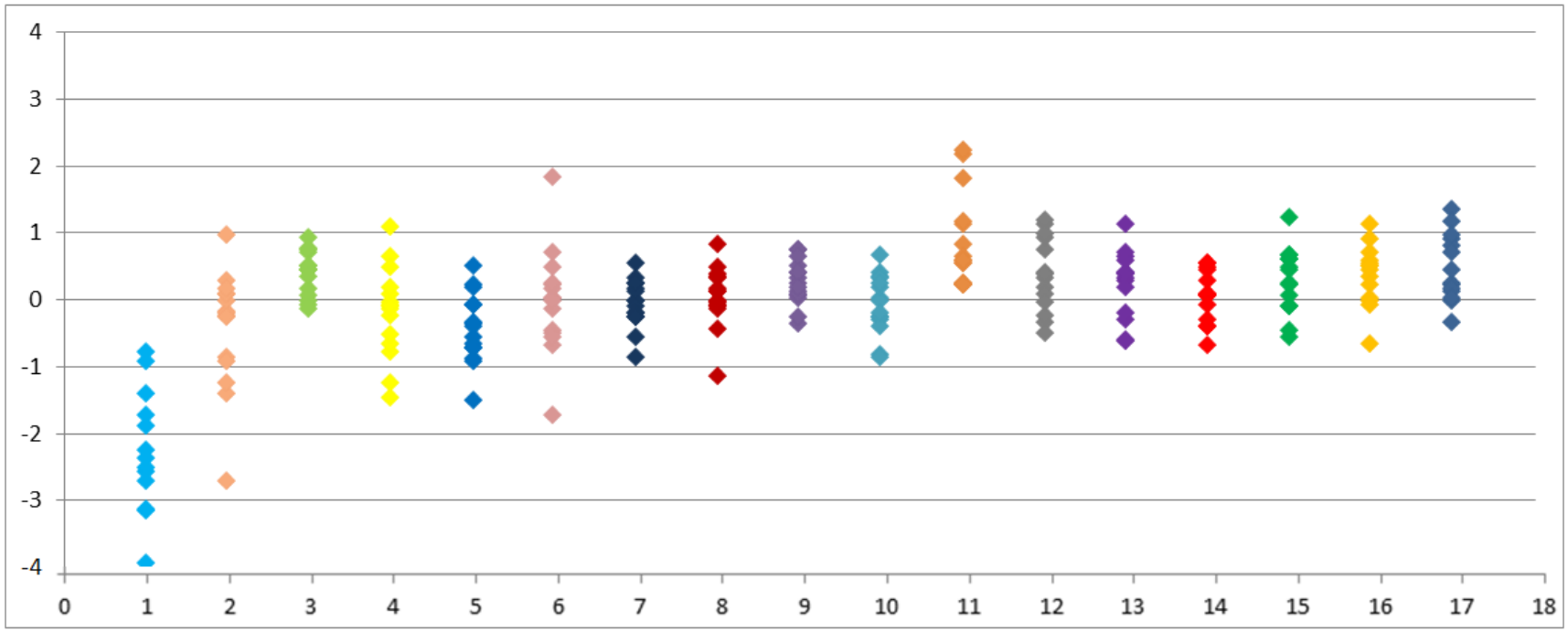

Figure 47. Difficulty levels of the items regarding each variable. The numbered line at the bottom shows the item numbers from 1 to 17. 1 is the item in Activity 1. 2-11 are the items in Activity 2 respectively. 12-17 are the first six items in Activity 3 respectively. The vertical axis represents the latent trait continuum from -4 to 4 logits. Each colored spot represents one variable measured by the item.