

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Philosophy and Methodology of Psychological Measurement for Social-Emotional Skills

Permalink

<https://escholarship.org/uc/item/3kq7v1rr>

Author

Suksiri, Weeraphat

Publication Date

2020

Peer reviewed|Thesis/dissertation

Philosophy and Methodology of Psychological Measurement for Social-Emotional Skills

by

Weeraphat Suksiri

A dissertation submitted in partial satisfaction of the

requirement for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

Of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair

Professor Sophia Rabe-Hesketh

Professor Lexin Li

Fall 2020

Abstract

Philosophy and Methodology of Psychological Measurement for Social-Emotional Skills

by

Weeraphat Suksiri

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

Operationalism, representationalism, and pragmatism are three major philosophies of measurement in psychology and education. These philosophies play important roles in designing and conducting measurements, as well as interpreting results. When measurement results are available, validity evidence plays important roles in supporting the plausibility of interpretation of measurement results and the appropriateness of use of measurement results. This paper focuses on discussing the nature and concepts of the philosophies and validity. Having this done, an integrative framework is described, called the construct modeling approach, in which the three philosophies and validity are situated to serve the complex reality in the fields. The assessment of Thai community attachment (TCA) is used to demonstrate how the framework works.

The assessment began with creating a developmental structure of community attachment, called the TCA construct map, to show five levels of the attribute, ranging from Receiving as the lowest level, to Complying, Accepting, Valuing, then Expanding as the highest level. Community attachment in Thai society can be described into three content domains, including (a) National, Religion, and King (NRK), (b) Thainess, and (c) Thai Citizenship. This yielded a three-dimensional structure of community attachment. Then, the developmental structure was situated in the three dimensions; that is, three construct maps of community attachment were developed—the NRK, Thainess, and Thai Citizenship construct maps.

Situational TCA items were created and assigned to grade 9th students in Thailand. Student scores or measurement results were intensively investigated in order to provide validity evidence to support the plausibility of score interpretation and the appropriateness of score use.

The results showed that when the construct modeling approach was used as the integrative framework, a well-designed measurement can be systematically established, and that measurement procedure empirically relates to the three underlying philosophies as well as its measurement results are supported with several validity evidence. Moreover, the integrative framework also provides useful information for the revision of the TCA assessment.

For my family, Mae Lek, Por Daeng, and Pee Joy

Table of Contents

List of Figures

List of Tables

Chapter

1	An Integrative Framework of Philosophy of Measurement and Validity Evidence.....	1
1.0	Introduction.....	1
1.1	Operationalism, Representationalism, and Pragmatism about Psychological Measurement.....	2
1.2	Validity Evidence in the 2014 Standards.....	4
1.3	Construct Modeling Approach as an Integrative Framework of Philosophy of Measurement and Validity Evidence.....	5
1.4	The Missing Components of Psychological Measurement of Social-Emotional Skills.....	8
1.5	The Affective Domain Taxonomy as a Structure of Social-Emotional Skills.....	10
1.6	Conclusion.....	13
2	An Application of the Integrative Framework to Thai Community Attachment Assessment.....	14
2.0	Introduction.....	14
2.1	Constructing Thai Community Attachment Construct Maps.....	15
2.2	Developing Situational Items and Scoring Guide.....	23
2.3	Establishing Measurement Models for Parameter Estimation.....	31
3	Results and Validity Evidence for the Thai Community Attachment Assessment..	39
3.0	Introduction.....	39
3.1	Response Processes Validity Evidence.....	39
3.2	Sample.....	39
3.3	Model Fit Comparison.....	40
3.4	Reliability Comparison.....	42
3.5	Standard Error of Measurement Analysis.....	42
3.6	Item Statistical Quality Evaluation.....	45
3.7	Item Calibration Validation.....	47
3.8	Item-Level Validation.....	55
3.9	Model Invariance Investigation and Measurement Revision.....	56
3.10	Score Interpretation.....	65
4	Conclusion.....	68
	References.....	69

List of Figures

1.1	A pragmatic-realist view of measurement.....	3
1.2	The relationship between validity evidence and measurement philosophy.....	5
1.3	Construct modeling approach.....	6
1.4	The relationship between the construct modeling approach, the measurement philosophy, and validity evidence.....	7
1.5	A potential form of the developmental structure.....	9
1.6	An example of dimensional structure.....	10
1.7	The levels of affective domain taxonomy.....	11
1.8	An example of the combination of dimensional and developmental structures.....	12
2.1	The levels and student's description of affective domain taxonomy.....	16
2.2	Scope of content domains of Thai community attachment.....	18
2.3	The NRK construct map.....	20
2.4	The Thainess construct map.....	21
2.5	The Thai Citizenship construct map.....	22
2.6	The graphical presentation of the definition of Thai community attachment.....	23
2.7	Examples of constructed-response TCA item.....	24
2.8	An examples of selected-response NRK item and response scores.....	26
2.9	An examples of selected-response Thainess item and response scores.....	27
2.10	An examples of selected-response Thai Citizenship item and response scores.....	28
2.11	The TCA items specification.....	29
2.12	The graphical presentation of the construct maps situated in dimensions and items.....	30
2.13	The graphical presentation the relationship between degree of attribute and item responses.....	31
2.14	Measurement models.....	32
3.1	Correlation matrix of student's TCA attributes.....	41
3.2	Graphs and descriptive statistics of standard error of measurement.....	44
3.3	The TCA Infit mean square graph.....	46
3.4	The Wright map of NRK dimension.....	48
3.5	The Wright map of Thainess dimension.....	49
3.6	The Wright map of Thai Citizenship dimension.....	50
3.7	The revised Wright map of NRK dimension.....	52
3.8	The revised Wright map of Thainess dimension.....	53
3.9	The revised Wright map of Thai Citizenship dimension.....	54
3.10	The TCA Infit mean square graph (total sample).....	57
3.11	Graphs and descriptive statistics of standard error of measurement (total sample).....	58
3.12	Correlation matrix of student's TCA attributes (total sample).....	59
3.13	The Wright map of NRK dimension (total sample).....	61
3.14	The Wright map of Thainess dimension (total sample).....	62
3.15	The Wright map of Thai Citizenship dimension (total sample).....	63
3.16	The TCA criterion zone.....	67

List of Tables

3.1	Model fit statistics comparison.....	40
3.2	EAP/PV reliability comparison.....	42
3.3	The mean locations trajectories of student attributes in each level of TCA.....	55
3.4	Comparison of variance-covariance matrices of the multidimensional model.....	56
3.5	EAP/PV reliability (total sample).....	57
3.6	The mean locations trajectories of student attributes in each level of TCA (total sample).....	64
3.7	Descriptive statistics of student attributes from the dimension alignment analysis.....	65
3.8	Item thresholds statistics.....	66

Chapter 1

An Integrative Framework of Philosophy of Measurement and Validity Evidence

1.0 Introduction

“...when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be.”
(Lord Kelvin cited in Thomson, 1889, p.73)

The statement above obviously considers measurement as an integral part of the scientific inquiry that involves a process of representing attributes¹ of objects² with quantity (de Gruijter & van der Kamp, 2008; Netemeyer et al., 2003; Tal, 2017). Measures and their psychometric properties do not depend only on the type of statistical measurement model undertaken and the context and goals of the measurement, but also rely heavily on the philosophical foundations of the measurement. In recent years, the philosophical foundations of measurement used in physical sciences and natural philosophy have gained attention after psychological researchers and psychometricians have tried to apply them to the field of psychology to improve measurement procedure in the field (Maul et al., 2016). This application helps psychological researchers and psychometricians gain understanding of the nature of psychological attributes and the logic of measurement, then leads them to designing an appropriate measure that hopefully provides valid measurement results according to the *Standards for Educational and Psychological Testing* (the 2014 Standards: AERA et al., 2014).

There have been three main influential measurement philosophies used in psychology: operationalism, representationalism, and pragmatism. When introducing these three philosophies to elaborate the logic of psychological measurement, the philosophies do not directly compete with each other to become the best appropriate one for the field. In fact, they can be integral parts and demonstrate different roles under the construct modeling approach (Wilson, 2005), and provide different types of evidence for arguing or supporting the validity of measurement results according to the 2014 Standards.

The well-known definition of measurement widely used in psychology by S.S.Stevens is the assignment of numbers to objects or events according to rules (Stevens, 1946) which make consistent and non-random assignment (Stevens, 1975). Stevens did not clearly discuss attributes

¹ In this paper, the term “attribute” refers to both “properties” and “relations” according to Maul et al. (2013). Sometimes a set of relevant attributes is used to refer to a construct or psychological trait (Junker, 1999; Netemeyer et al., 2003); however, the term attribute refers to a trait or a construct in this paper.

² The term “object” refers to “a system,” “process,” and “event” according to Tal (2017).

of objects in his definition, but it is important to consider that among numbers assigned to objects, different numbers convey different information about attributes.

Currently, even though there is no absolute conclusion among philosophers, psychology researchers, and psychometricians who hold different philosophical perspectives on the definition of measurement about the conditions that make measurement possible, and the attributes which can be measured, it is worth understanding the nature of measurement from different perspectives, especially when they are situated within the same measurement design framework and the same validation framework.

The main purpose of this paper is to demonstrate the relations between the three measurement philosophies: operationalism, representationalism, and pragmatism, and different types of validity evidence, according to the 2014 Standards, within the context of Thai community attachment using the construct modeling approach. The paper is structured to first provide an overview of the nature and concepts of the philosophies of measurement, then to review the concepts of validity evidence; finally, the paper will link these aspects to the construct modeling approach (Wilson, 2005) as an integrative framework by using an example (the assessment of student's sense of community).

1.1 Operationalism, Representationalism, and Pragmatism about Psychological Measurement

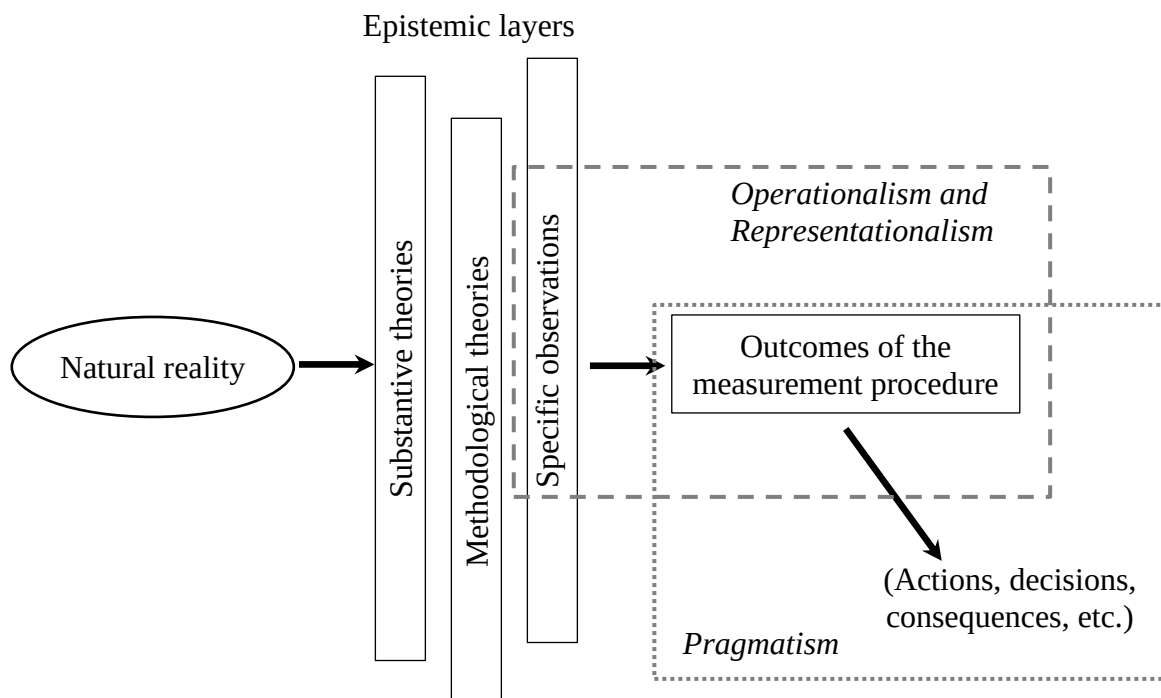
As stated previously, in principle, operationalism, representationalism, and pragmatism about measurement are consistent with each other since they highlight different and complementary aspects of measurement. Operationalism has a strong influence on psychology, more precisely on the behaviorism perspective in psychology. The philosophy provides psychological researchers and psychometricians a way to define psychological attributes by matching them with concrete operations. This philosophy is also associated with Stevens' view of measurement scale (Netemeyer et al., 2003; Tal, 2017), which views measurement as a set of operations that shape the meaning and/or regulate the use of a quantity term (Tal, 2017). It reflects a way that humans use to understand and describe the concept of objects being measured, called an operational definition, as well as making measurement possible. The concept of objects is, thus, synonymous with the corresponding set of operations (Bridgman, 1927). However, it is important to notice that when operations change, quantities or measurement results of an object are not always the same, and similarly, when the concept of an object is defined by different sets of operations, the quantities of the same object are also different.

Representationalism is currently one of the most influential philosophies of psychological measurement (Tal, 2017). Representationalism defines measurement as the construction of mappings from empirically relational structures into numerically or theoretically relational structures (Krantz et al., 1971; Maul et al, 2013). That means that there are two rational structures: one is the empirical structure which consists of a set of empirical or observable objects (e.g., student responses to an item) along with certain qualitative relations among them (e.g., classifying and ordering the responses), and another is the theoretical structure which consists of a set of numbers (e.g., 0, 1, 2, 3, and 4) and specific mathematical relations among them (e.g., $0 < 1 < 2 < 3 < 4$), and that these are connected. These relational structures reveal the standard order or sequence procedure underlying the measurement procedure of unobserved psychological attributes using observed relevant objects. This means psychological attributes can be conceptualized and structured using a standard sequence or order (Trendler, 2019) to reflect

the different degree of attributes being measured (Netemeyer et al., 2003). Connected to operationalism, representationalism clarifies “rules” that were mentioned in Stevens’s measurement definition (1946, 1951, 1968, 1975) as a known and consistent structure underlying objects before transforming to numbers.

Pragmatism views measurement as a set of procedures for interpreting the quantity of terms of objects and evaluating uses or practical consequences and applications of measurement (Maul et al., 2013). The main idea of pragmatism is whether measurement results can be usefully applied, rather than whether measured attributes are formally and ideally constructed out of their empirical relations (Maul et al., 2016). In psychological measurement, pragmatism is basically embedded in the measurement purposes. Every measurement has its own purpose (National Research Council [NRC], 2001); when psychological researchers or psychometricians design a measurement, then conduct it, and at the end of the measurement procedure, they interpret the measurement results, and they justify the results whether the results can be used to serve the purpose at the beginning.

Maul et al. (2013) distinguished the nature of operationalism, representationalism, and pragmatism. While operationalism and representationalism are motivated by the intuition that the preferred method of acquiring knowledge is through observation and experience, pragmatism is motivated by the intuition that theories are made useful via their links with practice. Thus, extreme proponents of operationalism and representationalism are overreaching in denying that psychological attributes exist independently of observations, while extreme advocates of pragmatism are overreaching in denying that measurement seeks to gain knowledge of a natural world.



Note. Modified from Maul et al. (2013)

Figure 1.1 A pragmatic-realist view of measurement

Figure 1.1 illustrates a framework that in psychological measurement, the operationalism, representationalism, and pragmatism overlap in the measurement results or outcomes, even though they encompass different and complementary aspects of measurement. In this framework, Maul et al. (2013) showed that psychological researchers and psychometricians understand the natural reality such as psychological attributes that exist in the real world, through substantive theories, methodological theories, and specific observations before they obtain measurement results or outcomes. When applying the logic of measurement in psychology to this framework, the operationalism and representationalism perspectives mainly focus on observations in acquiring measurement results, as shown in the dashed rectangle; while the pragmatism perspective mainly focuses on uses and practical applications of measurement results, as shown in the dotted rectangle.

1.2 Validity Evidence in the 2014 Standards

Psychological measurement is developed with a certain purpose(s) and its results are used to make decisions (Mellenbergh, 2011). This logic is consistent with pragmatism and the concept of validity in the 2014 Standards. Thus, the validity of measurement results has not been defined narrowly as the relation between an operational definition of a psychological attribute and items being used to measure according to operationalism, and merely as dimensions derived from either exploratory or confirmatory factor analysis according to representationalism.

The concept of validity has shifted to an argument-based approach in the 2014 Standards (Kane, 2006, 2013a, 2013b, 2016). Validity refers to interpretations and uses of measurement results and involves various sources of evidence that are used to evaluate the precision, plausibility, and appropriateness of the measurement. Because of this shift, it seems that operationalism, representationalism, and pragmatism are integrated into the same measurement results validation framework, and they provide different and complementary sources of validity evidence. When validity evidence is available, at least psychological researchers and psychometricians will know whether there is the need for refining the definition of the psychological attribute according to the operationalism perspective, whether items and measurement procedure have to be revised according to the representationalism perspective, and whether measurement results and underlying theory support the intended interpretation and specific uses according to the pragmatism perspective.

There are five types of validity evidence in the 2014 Standards. Evidence based on test content refers to how well the content in items is related to the psychological attribute intended to be measured. This type of evidence is closely tied to the alignment of the content of items to the definition of the attribute, which usually consists of “evaluating the correspondence between student learning standards and test content” (AERA et al., 2014, p. 15) in educational measurement settings. Clearly defining the attribute and the items based on the research literature and expert judgment are typical methods for gathering validity evidence of this type.

Evidence based on response processes, on the other hand, refers to whether respondents are responding to the items in the ways intended by psychological researchers and psychometricians as item developers. Thus, it is important to interview a sample of the intended respondent population to ensure that they are, in fact, using intended attributes to answer the items. Think-aloud interview is a method available to collect this useful information.

Evidence based on internal structure refers to “the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (AERA et al., 2014, p. 16). This means investigating whether the items measure at least approximately what they intend to measure. Items that measure additional, unintended constructs are said to have construct-irrelevant variance (Haladyna & Downing, 2004), while those that do not measure the intended constructs are called construct-deficient (AERA et al., 2014).

Evidence based on relations to other variables refers to the relationship of the items to other attributes or tests that may or may not be related to the attributes of interest. Evidence based on testing consequences refers to the use of the interpreted test scores as intended by the test developer to maximize the expected effects and to minimize the effects of unintended consequences.

Figure 1.2 illustrates that the five types of validity evidence relate to the three philosophies of measurement. Operationalism, which focuses on shaping the meaning of a psychological attribute with an operational definition, provides the evidence based on test content (TC) that indicates the relationship between the content of items and the content or definition of the psychological attribute. Representationalism, which focuses on the relationship between the theoretical structure and the empirically rational structures of the psychological attribute, provides the evidence based on internal structure (IS), response processes (RP), and relations to other variables (RV). Pragmatism, which focuses on the interpretation and use of the measurement results, provides the evidence based on testing consequences (CS).

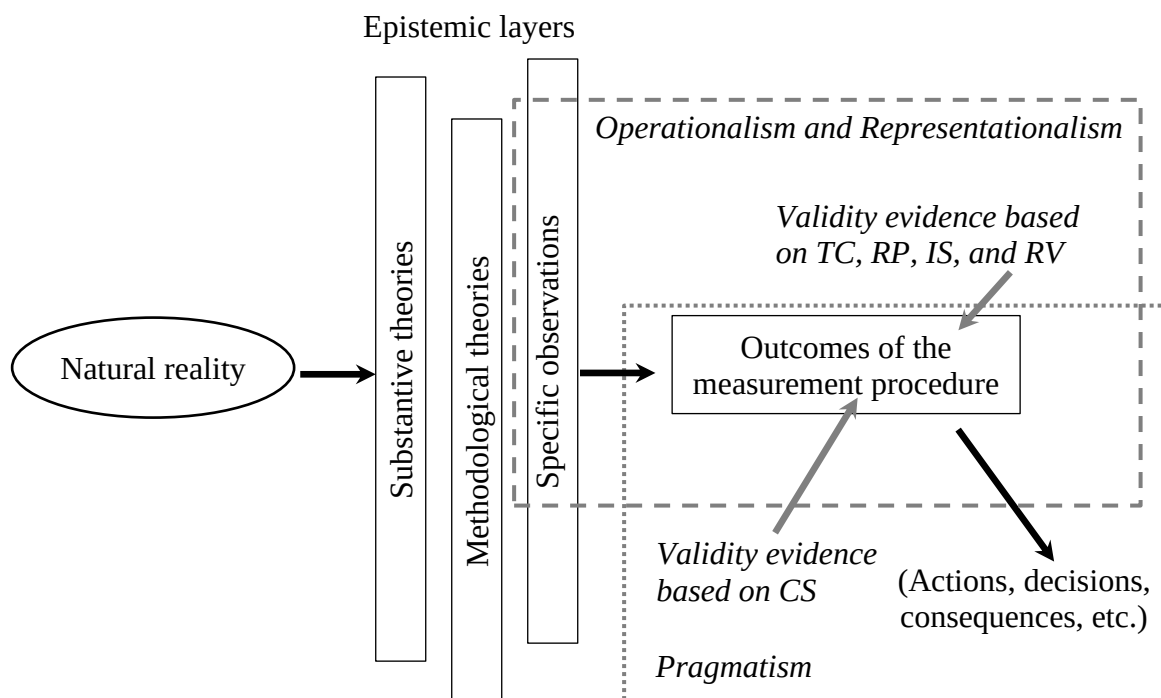


Figure 1.2 The relationship between validity evidence and measurement philosophy

1.3 Construct Modeling Approach as an Integrative Framework of Philosophy of Measurement and Validity Evidence

The construct modeling approach (Wilson, 2005) is an item response modeling approach for test or research instrument construction and development; and more recently it has also been used as a framework for designing measurement in the social sciences, especially in psychology and education. The approach consists of four components or building blocks starting with the construct map, item design, outcome space, and measurement model, respectively, as shown in Figure 1.3

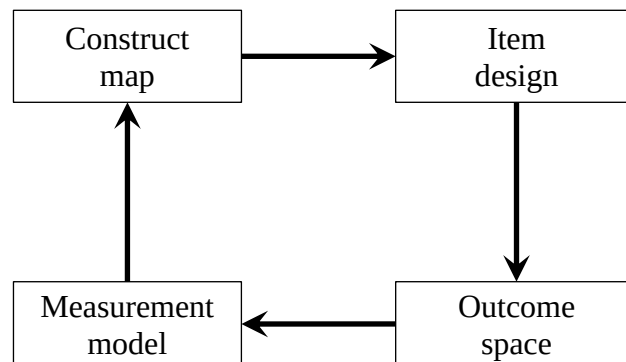


Figure 1.3 *Construct modeling approach*

The construct map is a model or structure of what is being measured such as psychological attributes. Every psychological measurement should begin by making sure what being measured really is. It represents the theory expressed as the multiple levels of the attribute that reflect characteristics of persons who possess different attributes in different amount of the attribute in degrees from low to high levels (Wilson, 2009). The construct map also includes a description of a typical person for each level, as shown in the next section, that plays an important role in defining the attribute and allows measurement results to be interpreted. Thus, the role of the construct map in terms of operationalism and representationalism is to match the theoretical structure of a psychological attribute with its operational definition.

The item design is the second component of the approach. It does not just refer to an item format, such as multiple-choice or open-ended formats, or just an item writing process. The term “design” used here has two distinct (but relevant) meanings (Maxwell, 2013; Wilson, 2005). The first meaning focuses on the procedure or activities involved in designing an item; that is, design is used as a verb. Another meaning focuses on product or result of designing; that is, design is used as a noun. Thus, item design is a process that we need to think about as we design and create items that most effectively elicit evidence or information from intended respondents who possess that psychological attribute, and the information should reflect different levels of the psychological attribute defined in the construct map. Thus, when respondents respond to the items, item design is a tool or product that defines the attribute with the items and responses according to the operational perspective and structures the response to be consistent with the theoretical structure or the construct map according to the representational perspective so that the

measurement is the most useful, informative, and interpretable with respect to the measurement purpose.

The outcome space describes in detail the qualitatively different categories of responses elicited by the items and maps these categories onto the level of the construct map. The outcome space is also a tool that facilitates the response scoring process so that each response category is assigned a unique code and score, and this corresponds to the levels of the construct map. For example, if response category B indicates a higher level of processing the attribute than response category A, that means B should be assigned a category number greater than the category number assigned to A. The ordering of scores, response categories, and levels of the construct, thus, must be made consistent and reflect the practical concept of the representation.

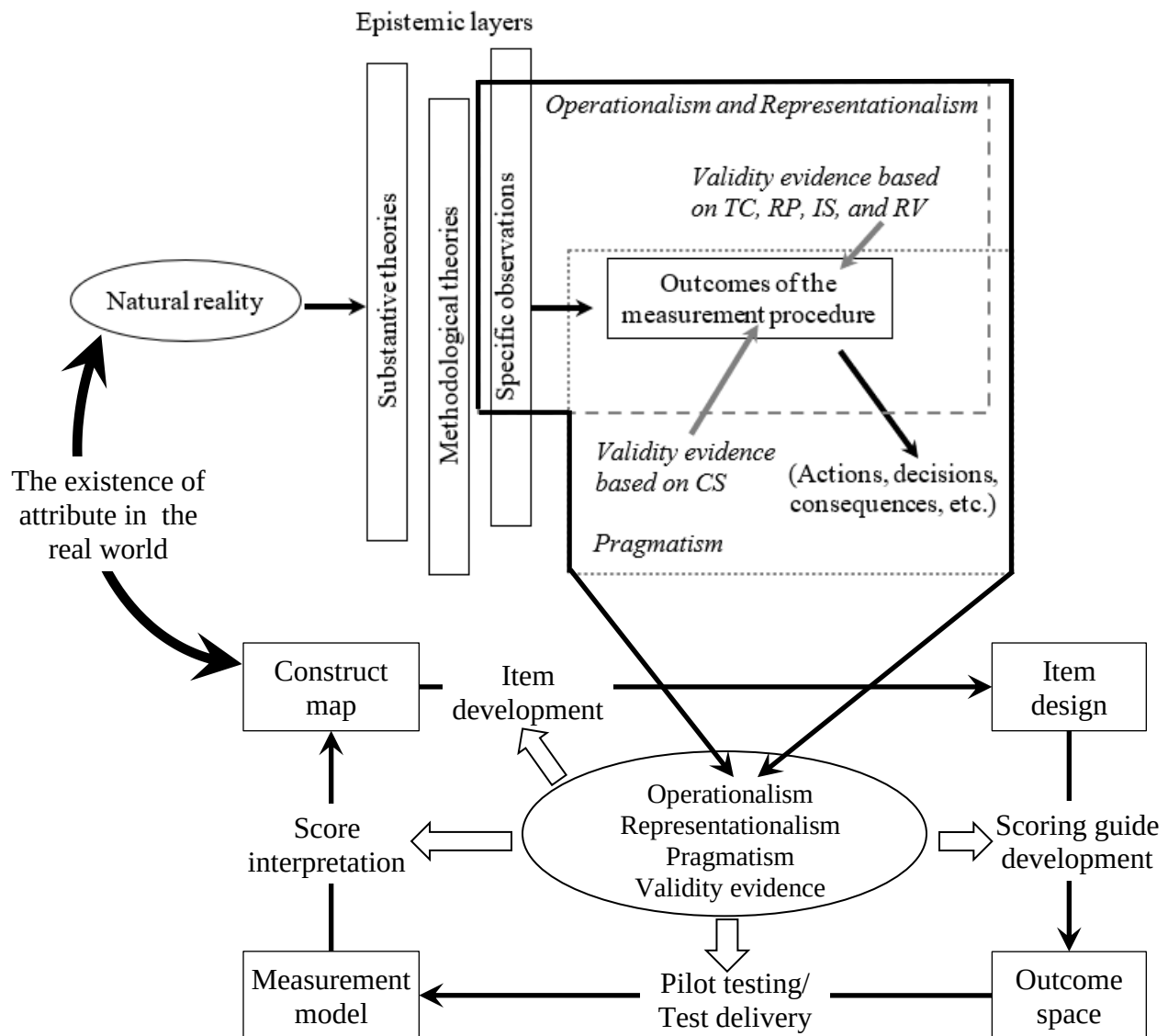


Figure 1.4 The relationship between the construct modeling approach, the measurement philosophy, and validity evidence

The measurement model is the last component of the approach. The chosen measurement model is the Rasch model (Rasch, 1960/1980) and its associated models' family. This model is chosen because it shapes the measurement design from the previous steps suitable for use in the construct mapping formulation. The Rasch model parameterizes the persons and the items in a way that psychological researchers and psychometricians are able to interpret in terms of the psychological attribute. The measurement results are presented in a visual form of the empirical construct map, called the Wright map. The Wright map is the center of interpretation and use of the measurement results. That means the measurement results and the Wright map reflect the idea of both the representationalism and pragmatism embedded in the construct modeling approach.

Figure 1.4 shows that the three measurement philosophies and validity evidence can be integrated into the construct modeling approach, at the center of measurement. The procedures in the construct map, item design, and outcome space steps mainly relate to the operational and representational perspectives and provides validity evidence based on test content, response processes, internal structure, and relations to other variables, while the measurement model step mainly relates to the representational and pragmatical perspectives and provides validity evidence based on testing consequences in a way that psychological researchers and psychometricians can interpret and use to make practical applications of measurement results.

1.4 The Missing Components of Psychological Measurement of Social-Emotional Skills

The construct map is the most fundamental component of the construct modeling approach as shown in Figure 1.4. It connects the three measurement philosophies and the validity evidence to measurement results and the reality. When an attribute is well-defined via the construct map, a set of items can be developed to elicit the intended information (as the item development procedure) and responses to the items can be scored according to the levels of the attribute in the construct map (as the scoring guide development procedure and the test delivery procedure). A measurement model is used to estimate psychometric parameters about the items; then the measurement results are interpreted according to the construct map (as the score interpretation procedure). Meanwhile, the measurement is conducted and information from every step can be used to support the claims being made from the measurement. However, similarly to educational measurement (Brown & Wilson, 2011; Mislevy, 1996; NRC, 2001), many psychological measurements lack underlying structures to define psychological attributes. This reflects an incomplete procedure in designing the measurements. That is the measurement results tend to fail to provide new understanding about psychological measurements (Keeves & Masters, 1999).

Social-emotional skills have played an important role in supporting the well-being of individuals and nations. They help individuals face the challenges of the modern world and help policymakers, teachers, parents, and communities to prepare children to be ready for achieving their diverse life goals in the future (OECD, 2015). These attributes are broadly defined as individual characteristics that productively drive people in achieving goals, working with others, and managing emotions in everyday life situations and in all stages of life. The skills, thus, involve individual capacities that (i) are manifested in consistent patterns of thoughts, feelings, and behaviors, (ii) can be developed through learning experiences, and (iii) influence important social and economic outcomes throughout the individual's life.

The framework usually used to classify the social-emotional skills is the Big Five personality taxonomy (OECD, 2015). This framework distinguishes five broad personality dimensions: extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience (Digman, 1990; Goldberg, 1992; John & Srivastava, 1999; McCare & John, 1992). However, when using this framework in assessing an individual's personality, the five dimensions need to be narrowed and specified in order to understand the individual's social-emotional skills development and interpret consistent patterns of the responses across situations.

Measuring the social-emotional skills requires a well-designed approach because the attributes cannot be observed directly. In the literature on personality measurement, psychological researchers and psychometricians place great emphasis on defining the meaning and scope or dimensional structures of the attributes according to operationalism (Kyllonen et al., 2005; Lipnevich et al., 2013) but devote little attention to progression structures embedded in the attributes according to the representationalism and pragmatism underlying the measurement designs. As described earlier, these missing developmental structures and measurement philosophies are considerably important because they bridge two gaps in the measurement procedure: one between the psychological attributes of individuals and measurable behaviors according to the representational perspective, and another between the measurement design procedure and interpretation of measurement results according to the pragmatism perspective. To measure such skills plausibly and appropriately, I would like to develop a measurement that includes the two missing components.

Every measurement is designed to meet intended purposes and to be used in a specific context (NRC, 2001). All measurement, however, share three key components: (i) a structure or model of how individuals represent their attributes, (ii) items or situations that allow one to observe individuals' behaviors, and (iii) an interpretation method for drawing inferences from the obtained evidence (Brown & Wilson, 2011). In every psychological measurement, a structure of how to represent attributes helps psychological researchers and psychometricians understand how individuals present, organize, and process attributes, and how they interact with society. Individuals with different degrees of attributes can be located on a developmental continuum structure according to representationalism. Figure 1.5 shows a simple theoretical structure for an attribute that can be conceptualized as a developmental continuum structure, running from a less or low level at the bottom to a more or high level at the top of the up arrow. Individuals with different degrees of an attribute will be mapped to the line to display their developmental structure.

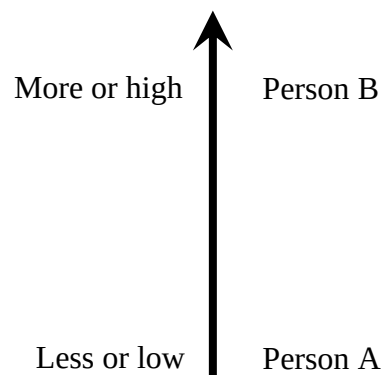


Figure 1.5 *A potential form of the developmental structure*

The idea of developmental structure in psychological measurement design was promoted in a 2001 NRC report (NRC, 2001). However, there are only a few studies applying this notion to social-emotional domains. Most attributes relevant to social-emotional skills are measured using self-report items with Likert-type scales, ranging from strongly disagree to strongly agree. These measurement scales, however, are not explicit structures of the social-emotional skills, and hence it is relatively difficult to interpret the meaning of increasing and decreasing amounts of the attributes. In particular, these provide limited information about stages or progression in thoughts, feelings, and behaviors of individuals. Thus, that means many current measures of social-emotional skills tend to not integrate well the measurement philosophies of operationalism, representationalism, and pragmatism into their measurement designs and procedure to gain new understanding about the social-emotional skills, the measurements tend to lack strong validity evidence to support the plausibility of interpretation of measurement results and the appropriateness of use of measurement results.

1.5 The Affective Domain Taxonomy as a Structure of Social-Emotional Skills

I assume that the structures of social-emotional skills are dimensional and developmental structures. The dimensional structure involves distinguishing key dimensions of the attributes according to certain criteria (e.g., content or concept). For example, if the social-emotional skills are conceptualized as the Big Five, the dimensional structure would comprise five broad dimensions, including open-mindedness, task performance, emotional regulation, engaging with others, and collaboration (OECD, 2015). Meanwhile, if psychological researchers and psychometricians focus on one of the five dimensions such as collaboration, the dimensional structure would consist of empathy, trust, and cooperation subdimensions (OECD, 2015) as shown in Figure 1.6. Note that panel (a) and (b) in Figure 1.6 are mathematically identical (Edwards et al., 2015).

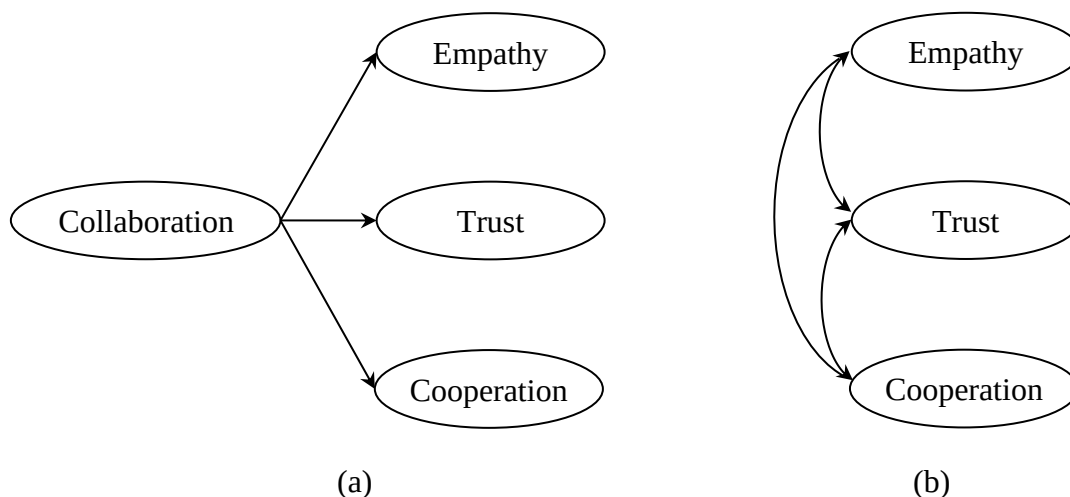


Figure 1.6 *An example of dimensional structure*

The developmental structure refers to a model of quantities or levels of presence of attribute. The structure should be based on a monotonic increasing property, corresponding to the development of social-emotional skills. I agree with Krathwohl et al. (1964), Mischel and Shoda (2008), and Rosenberg and Hovland (1960) that feeling relates to and interacts with thinking or cognition and behavior in responding to situations. Thus, the measurement of social-emotional skills should also follow these relationships.

I will apply the affective domain taxonomy (Krathwohl et al., 1964) to the measurement of social-emotional skills. The taxonomy refers to a framework of describing and referring to individual characteristics in action or response to different situations. The taxonomy classifies a range of individual responses, including knowing about something, solving problems, revealing an interest in some types of experience, and exhibiting behaviors to situations. Thus, this taxonomy can be used to describe an individual's attributes in different degrees of possessing these social-emotional skills.

The construct structure positions different affective responses on a continuum to show certain types of learning in a learning theory, which are relevant to the principles of “simple to complex” (Krathwohl et al., 1964). The affective taxonomy greatly emphasizes feelings or emotions. It is consistent with the idea that individuals cannot feel without thinking, and cannot act without feeling and thinking. Thus, the order of response patterns in the taxonomy starts from simple thinking, goes to complex feeling, then to consistent behaving.

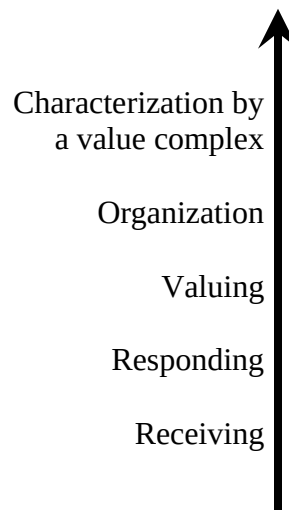


Figure 1.7 *The levels of affective domain taxonomy*

Figure 1.7 shows a structure for the affective domain taxonomy. The taxonomy has five levels, beginning with Receiving, as the lowest level at the bottom of the up arrow, with individuals being aware of situations in initiating the affective behavior. At the next level, Responding, the individuals are perceived as responding regularly to situations. Thus, the individuals have differentiated the situations and begin to seek them out and to attach emotional significance and value to them.

The next level is Valuing which describes increasing internalization as the individual's behavior is sufficiently consistent that they hold a value. When the individuals successfully internalize values, they encounter situations where more than one value is relevant. This necessitates organizing the values into a system. So, this next level is Organization. Finally, at Characterization, shown as the highest level at the top of the up arrow, the internalization and the organization processes reach a point where the individuals respond consistently to situations with an interrelated set of values.

I will use the taxonomy as a developmental structure of social-emotional skills and situate it into a dimensional structure as a way of integrating the operationalism and representationalism. Considering Figures 1.6 and 1.7 where two structures underlie the collaboration attribute, the developmental structure can be embedded in the dimensional structure as shown in Figure 1.8. That means each dimension has its own set of different levels of presence of the attribute for individuals. Figure 1.8 shows a 3D visualization of the structures of the collaboration construct. The dimensional structure of empathy (E), trust (T), and cooperation (C) is displayed in the XY plane; and the developmental structure is on the Z-axis. The up arrows still preserve their meaning in describing the different levels of presence of each of the dimensions.

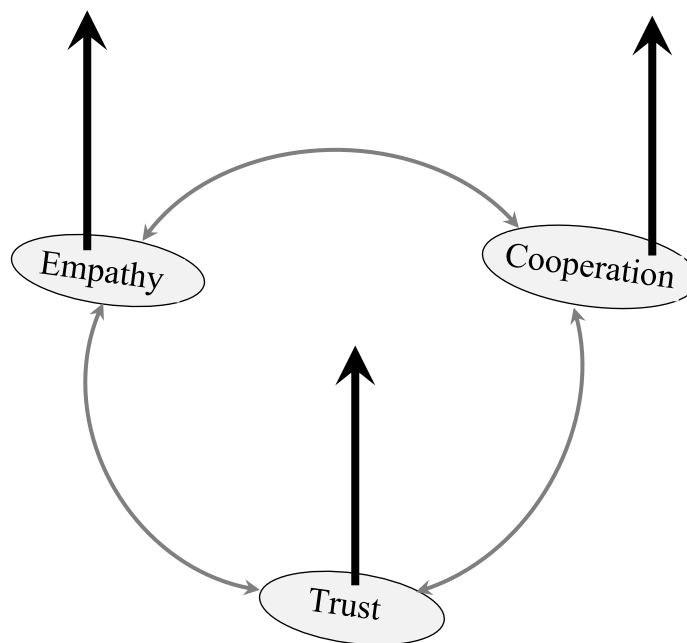


Figure 1.8 *An example of the combination of dimensional and developmental structures*

1.6 Conclusion

Operationalism, representationalism, and pragmatism are three major philosophies of measurement in psychology and education. These philosophies play important roles in designing and conducting measurements, as well as interpreting results. When measurement results are available, validity evidence plays important roles in supporting the plausibility of interpretation of measurement results and the appropriateness of use of measurement results. This paper focuses on discussing the nature and concepts of the philosophies and validity. Having this done, an integrative framework is proposed, called the construct modeling approach, in which the three philosophies and validity are situated to serve the complex reality in the fields. The assessment of Thai community attachment is used to demonstrate how the framework works.

Chapter 2

An Application of the Integrative Framework to Thai Community Attachment Assessment

2.0 Introduction

In order to demonstrate the application of the three philosophies of measurement and the validity evidence of measurement results that are situated within the same measurement to a psychological measurement of social-emotional skills, the measurement of community attachment in Thailand for grade 9 students is used. The study aims to develop a tool for assessing the attachment to community, which is currently raised as a core value of Thai society. The study also aims to develop a reasonable guideline for describing the individual differences about community attachment. The overview of methods and design are described as follows.

(i) The Thai community attachment (TCA) construct map was created based on a literature review. The construct map consisted of five levels of possessing the community attachment, ranging from *Receiving* as the lowest level, to *Complying*, *Accepting*, *Valuing*, then *Expanding* as the highest level. These yielded a five-levels developmental structure of community attachment attribute. Community attachment in Thai society can be categorized into three content domains—(a) National, Religion, and King (NRK), (b) Thainess, and (c) Thai Citizenship. These yielded a three-dimensional structure of community attachment, and then the developmental structure was situated in the three dimensions. That is, three construct maps of community attachment were developed—the NRK construct map, Thainess construct map, and Thai Citizenship construct map.

(ii) Ninety situational items in constructed-response format were created to measure the TCA according to the three construct maps. All situations used in the items were derived from a focus group with teachers about everyday events and activities occurring in communities in which students live. Each item was designed to measure only one dimension of the TCA contents, but it can measure more than one level of the TCA levels, depending on the nature of the situation used in that item. All items asked about how a student thinks and feels and how they decide to behave in each situation.

(iii) The items were piloted with 283 grade 9th students, then actual responses to the items were analyzed to obtain response patterns of thinking-feeling-doing combinations to create options for the items. There were 70 items that were transformed to selected-response format. The options of each item were classified into the five levels. It was not necessary to have options measure all five levels for each item. Individual items had a different number of options and some options within the same item could measure the same level of possessing community attachment.

(iv) The selected-response items were tested with the intended 980 grade 9th students. Their response data was divided into two datasets—a calibration dataset for estimating psychometric parameters and another validation dataset for evaluating the parameters and results. Psychometric parameters were estimated using the multidimensional random coefficients multinomial logit model (Adams et al., 1997). Validity evidence according to the *Standards for*

Educational and Psychological Testing (AERA et al., 2014) were provided to support the plausibility of interpreting the measurement results and the appropriateness of use of the measurement results. Moreover, the three philosophies of measurement were mentioned and discussed throughout the study in order to demonstrate how they were integrated into the procedures and results of the measurement.

2.1 Constructing Thai Community Attachment Construct Maps

As described in the early section, representationalism plays an important role in defining an attribute being measured as a way of mapping an empirical relation structure into a theoretical relation structure. That means, a theoretical structure of any attribute needs to be established at the early stage of the measurement procedure. In the TCA measurement, the affective domain taxonomy (Krathwohl et al., 1964) was used as an initial developmental structure of the TCA. The taxonomy consisted of five main categories—*Receiving*, *Responding*, *Valuing*, *Organization*, and *Characterization*, and the description of students, ordered from low to high levels along a continuum of internalization, as indicated by the up arrow in Figure 2.1. Note that all subcategories shown in the original taxonomy, such as Acquiescence in responding, Willingness to respond, and Satisfaction in response within the same Responding category, are not displayed in the figure because there is not sufficient evidence to support the existence of the ordering of those subcategories (Martin & Biggs, 1986), so in this measurement all subcategories are treated equally and refer to different pieces of affective information.

Figure 2.1 shows the affective domain taxonomy in construct map style. Each level has its own description of students' attribute. The levels show how a student develops their characteristics about an attribute. The taxonomy has several benefits to TCA assessment. First, the taxonomy is a classification structure of internalization that describes students' attributes in terms of the interaction between thoughts and feelings to relevant behaviors (Findley & Woodruff, 1964; Krathwohl et al., 1964; Krathwohl, 1968; Moseley et al., 2005). The internalization relates to both thinking and feeling (Anderson & Bourke, 2000; Martin & Biggs, 1986; Messick, 1979), and drives individuals who possess the same attribute to behave differently (Larsen & Buss, 2010; Mischel & Shoda, 1995, 2008), then their behaviors differentiate them from others (McCoach et al., 2013). This explanation corresponds to the nature of community attachment that consists of cognitive, affective, and behavioral components to allow individuals to acquire, maintain, and express community attachment to serve their personal needs (Bar-Tal & Staub, 1997; Primoratz, 2017) and corresponds to the concept representational perspective.

Another benefit is that the taxonomy primarily focuses on the structure of internalization for positive social values (Krathwohl et al., 1964), such as cooperation and empathy rather than aversion and aggressiveness. This is the way that all educational objectives or learning standards are typically posited concerning desired social values and social-emotional skills. The taxonomy describes students' motivation and determination of social values, from low to high levels of possessing those values; that means, students' feelings and emotions can be described from neutral to mild, then to a positive stage, rather than from negative to positive (Findley & Woodruff, 1964; Krathwohl et al., 1964). This benefit supports the nature of TCA as a positive social value that teachers and parents take responsibility to inculcate the attribute to students. It also supports the intended interpretation of measurement results of the community attachment in

community settings. Students with a low degree of community attachment does not mean that they have none of the attribute; instead, it means that they show less learning of the desired attribute, and that opens the window to improve their learning process and outcome of community attachment.

Level	Description of student
Characterization by a value or value complex	Students adapt or apply a value system to control their behaviors in desired and undesired phenomena. They act consistently and persistently in accordance with the values they have internalized. They integrate beliefs, ideas, and attitudes to form characteristics.
Organization	Students conceptualize values when they encounter phenomena in which more than one value is relevant. They internalize and organize the values into a consistent system, determine interrelationship between the values to those that they already hold or to new ones that they are coming to hold, and establish dominant values and pervasive values to form the ordered relationship.
Valuing	Students use their own criterion of worth to assess a value of phenomenon to them. When a student is holding a value, they are both sufficiently consistent that others can identify the value, and sufficiently committed to the value guiding their behaviors, from they are willing to be, want it, deepen their involvement with it, and they try to convince others about their cause.
Responding	Students are sufficiently motivated and interested, then actively attending, and committing themselves to phenomena. They respond to the phenomenon with acquiescence or compliance, or even with willingness to gain more understanding about the phenomena with a learning-by-doing process. They can find themselves engaging in the phenomena with satisfaction and they can respond to the phenomena conforming to social norms, without full acceptance of the necessity for responding and their internal aspects of the behavior are quite different.
Receiving	Students are sensitized to the existence of certain phenomena; this is, they are willing to receive or to attend to the phenomena. They are oriented to learn what others intend that they will, and then they bring to each phenomenon a point of view which may facilitate or hinder their recognition of the phenomena to which the others are trying to sensitize them. They simply notice phenomena without discrimination and interest. They know phenomena superficially as directly perceived and involve the phenomena with neutrality.

Figure 2.1 *The levels and student's description of affective domain taxonomy*

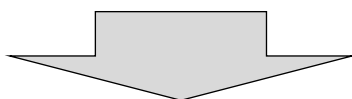
The final benefit is that the taxonomy is an internalization framework. It can be adapted to inculcating several traits and describes a complexity in trait change consistently when students learn to behave appropriately to establish and maintain a satisfying relationship with others, even if they have not completely accepted certain values (Krathwohl et al., 1964; Martin & Biggs, 1986). The taxonomy is flexible enough to describe stages of internalization of any attribute, where students at different levels have different attributes and students do not necessarily begin their internalization at the lowest level of the taxonomy (Findley & Woodruff, 1964). Students are able to enter the taxonomy at any level depending on their learning experience. With this benefit, the taxonomy can be modified to create a developmental structure of TCA, where the community attachment increases from low to high levels of possessing the attribute in several situations. When students at any level of the taxonomy are motivated by the situations, they can start their learning from a neutral mode and progress to a higher level (Findley & Woodruff, 1964; Krathwohl et al., 1964; Roberts, 2009); that means, student's learning process in the internalization of TCA occurs.

Another structure of TCA is a dimensional structure. The TCA's dimensional structure refers to three content domains obtained from a content analysis of four government documents—the twelve core values for Thai people (Ministry of Education of Thailand, 2014a), the eight desirable characteristics for the Thai students stated in the 2008 Basic Education Curriculum (Ministry of Education of Thailand, 2008), the five strands of Thai Citizenship education (Ministry of Education of Thailand, 2014b), and the ten duties of Thai people as described by the 2017 Constitution of the Kingdom of Thailand (Secretariat of the House of Representatives of Thailand, 2017) as shown briefly in Figure 2.2. This structure is a typical way to define an attribute in psychological measurement according to the operational perspective.

Figure 2.2 shows that key contents in the four sources are overlapping in terms of the expectation of attachment to Thai society. After conducting the content analysis of these documents and concepts, three content domains were found to be suitable for grade 9 students, including (a) the nation, religion, and the king (NRK), (b) Thainess, and (c) Thai Citizenship. Each dimension has three pieces of information that shape the concepts of the dimension. The nation, religion and the king dimension consists of (a1) the establishment of the nation, history, independence, and heroes, (a2) the fundamental principles of religions, morality, and doctrines, and (a3) the King's and royal's duties and the King's guidance and statement. The Thainess dimension consists of (b1) language, (b2) indigenous wisdom and knowledge, art and culture, tradition, and manners, and (b3) features of characters, including generosity, kindness, responsibility, and harmony. The Thai Citizenship dimension consists of (c1) social rules and norm, laws and regulations, and duties to oneself, family, school, community, and the nation, (c2) application of a Sufficiency Economy philosophy³, (c3) legal and human rights, respect for others' rights, and to exercises these rights with responsibility.

³ Sufficiency Economy is a philosophy that stresses the middle path of living with sustainability. It starts from the level of families to communities and to the nation in terms of development and administration, so as to modernize in line with the forces of globalization. Sufficiency means moderation, reasonableness, and the need for self-immunity to protect from impacts arising from internal and external change (Office of the National Economic and Social Development Board, 2007).

The Ten Duties
(1) Protecting and upholding the nation, religion, the king, and the democratic regime of government with the king as head of state, (2) Defending the country, protecting and upholding honor and interests of the nation and public domain of state, as well as cooperating in preventing and mitigating disasters, (3) Strictly observing the law, (4) Enrolling in compulsory education, (5) Serving in armed forces as provided by law, (6) Respecting and not violating the rights and liberties of other persons and not committing any act which may cause disharmony or hatred in society, (7) Freely exercising his or her right to vote in an election or referendum, (8) Cooperating and supporting the conservation and protection of the environment, natural resources, biodiversity, and cultural heritage, (9) Paying taxes and duties as prescribed by law, and (10) Not participating in or support all forms of dishonest act and wrongful conduct
The Twelve Core Values
(1) Upholding the three main pillars of the country: the nation, the religions, and the monarchy, (2) Having honesty, sacrifice, and patience with positive attitude for the interest of the public, (3) Expressing reverent gratitude to parents, guardians, and teachers, (4) Rendering diligence in learning and education, (5) Treasuring Thai traditions and cultures, (6) Practicing morality, integrity, considerateness, generosity, and sharing, (7) Understanding and learning true democratic ideals with the king as head of state, (8) Maintaining discipline, respectfulness of laws and the elderly, (9) Being conscious and mindful of action in line with the king's royal statements, (10) Applying the king's sufficiency economy, saving money for time of need, being moderate with surplus for sharing or expansion of business while having good immunity, (11) Keeping physical and mental strength, unyielding to the evil power or desires, having sense of shame over guilt and sins in accordance with the religious principles, and (12) Putting the public and national interest before one's own
The Eight Desirable Characteristics
(1) Love of nation, religion, and king, (2) Honesty and integrity, (3) Self-discipline, (4) Avidity for learning, (5) Observance of principles of sufficiency economy philosophy (Thai development approach to sustainability) in one's way of life, (6) Dedication and commitment to work, (7) Cherishing Thainess, and (8) Public-mindedness
The Five Strands of Thai Citizenship Education
(1) The nature of Thai people including Thai culture and Thai characters, (2) Attachment to the nation, religion, and the royal family, (3) Good citizenship, (4) Harmony and conformity, and (5) Self-discipline



Thai Community Attachment		
Nation, Religion, and the King	Thainess	Thai Citizenship
(a1) Establishment of the nation, history, independence, and heroes, (a2) Fundamental principles of religions, morality, and doctrines, and (a3) King's and royal's duties and King's guidance and statements	(b1) Language, (b2) Indigenous wisdom and knowledge, art and culture, tradition, and manners, and (b3) Features of characters including generosity, kindness, responsibility, and harmony	(c1) Social rules and norm, laws and regulations, and duties to oneself, family, school, community, and the nation, (c2) Application of sufficiency economy philosophy, and (c3) Legal and human rights, respect to others' rights, and exercise rights with responsibility

Figure 2.2 *Scope of content domains of Thai community attachment*

The nine concepts define the scope of TCA in a way that is consistent with operationalism. However, the scope of the TCA contents does not clearly define what community attachment really is and what levels of possessing community attachment really are. In the literature review, there is broad agreement on the meaning of community attachment as patriotism. For instance, Bar-Tal and Staub (1997) defined patriotism as a deeply felt affective attachment to communities in which they reside; similarly for Kosterman and Feshbach (1989) who defined attachment as the degree of love for and pride in one's community.

In Bar-Tal and Staub's view (1997), the meaning of attachment relates closely to patriotism that is a binding affection between a person and their community and can be viewed from three perspectives. From a psychological perspective, patriotism has cognitive, affective, and behavioral components. Individuals acquire it, maintain it, and express it in the service of their personal needs. From a social perspective, although patriotism is held by individuals, it develops only in groups, communities, or societies that provide meaning to and characterize the groups. From a political perspective, patriotism is a powerful tool for organizing, mobilizing, and gaining the support of group members. However, in this study, only the psychological and social perspectives are used to define the community attachment and the measurement purpose has no part relevant to any political issue; thus, in order to avoid any possible misunderstanding, the attribute of interest is called "community attachment". Moreover, the two perspectives are useful and consistent with the scope of TCA because in a school environment, Thai students have learnt about community attachment and social norms in cognitive, affective, and behavioral aspects.

In order to support the definition and scope of community attachment according to operationalism, community attachment is defined as relational structures which represent ordered levels or degree of attachment according to representationalism. With the idea of construct mapping in the construct modeling approach (Wilson, 2005), the TCA construct maps were mainly created based on the affective domain taxonomy or levels of internalization. The taxonomy corresponds to both the psychological perspective and the social perspective, as well as relating to the cognitive and behavioral domains that have influence on the Thai education system. It also fits with the learning process in the school environment and is developed to handle primarily positive characteristics associated with educational objectives which are generally stated in teaching plans, school activities, and even curriculum. Thus, results obtained for the affective domain taxonomy can be interpreted as less likely to more likely to achieve an affective educational objective.

As described above, therefore, the affective domain taxonomy was modified to be suitable for Thai educational purposes and the concepts of TCA; that is, the NRK, Thainess, and Thai Citizenship construct maps have the same structure of possessing community attachment, but different descriptions of student attributes, depending on the content of the dimension, as shown in Figures 2.3 to 2.5, respectively. Each dimension has five levels, with Receiving as the lowest level to Complying, Accepting, Valuing, and Expanding at the top.

As I have tried to define the TCA in a way that is consistent with representationalism and operationalism measurement philosophies, the three construct maps of NRK, Thainess, and Thai Citizenship represent the theoretical rational structures of the psychological attribute and integrate the operational definitions into the structures as shown in Figure 2.6. At this stage of the assessment design, the representationalism and operationalism do not operate separately. They support each other to define the underlying nature of TCA. Moreover, it is important to notice that the definition of TCA logically corresponds to the pragmatism measurement

philosophy to serve the purposes of the assessment ensuring that the intended interpretation and use of measurement results can be made.

Level	Description of student
Expanding (EP)	Students are concerned about the benefits that others will earn from their behaviors about the NRK. Thus, they try to convince others to their cause and try to expand the value of NRK to others with various forms of thinking, feeling, and behaving.
Valuing (VL)	Students conceptualize values and use their own criterion of worth to assess values in the NRK phenomena to them. They are satisfied to take the value of NRK as part of their belief; thus, their behaviors are sufficiently consistent and stable in various situations, from desired to undesired situations, and from simple to complex situations. This means they are holding the value and concerns about the benefits that they will earn from behaving.
Accepting (AC)	Students are concerned about the worth of NRK and seek to understand how to behave appropriately. They are willing to gain more understanding about the NRK phenomena. Students accept the worth of NRK as part of their beliefs and behave based on their own ways and social norms.
Complying (CP)	Students are sufficiently motivated and interested, then actively attending, and committing themselves to the NRK phenomena. They respond to the phenomena with compliance; that is, they respond to the phenomena consistently with a group or people, social norms, or their teacher's suggestions without their own reasons or full acceptance of the necessity for responding.
Receiving (RC)	Students are sensitized to the existence of NRK phenomena; that is, they are willing to receive or to attend to them. Students are oriented to learn what teachers intend that they will, and then they bring to each phenomena a point of view which may facilitate or hinder their recognition of the NRK to which the teachers are trying to sensitize them. Students perceived the phenomena without conscious or purposive effort.

Note. The term phenomena refers to situations, stimuli, activities, objects, persons, objects, or state of affairs.

Figure 2.3 *The NRK construct map*

Level	Description of student
Expanding (EP)	Students are concerned about the benefits that others will earn from their behaviors about Thainess. Thus, they try to convince others to their cause and try to expand the value of Thainess to others with various forms of thinking, feeling, and behaving.
Valuing (VL)	Students conceptualize values and use their own criterion of worth to assess values in the Thainess phenomena to them. They are satisfied to take the value of Thainess as part of their belief; thus, their behaviors are sufficiently consistent and stable in various situations, from desired to undesired situations, and from simple to complex situations. This means they are holding the value and concerns about the benefits that they will earn from behaving.
Accepting (AC)	Students are concerned about the worth of Thainess and seek to understand how to behave appropriately. They are willing to gain more understanding about the Thainess phenomena. Students accept the worth of Thainess as part of their beliefs and behave based on their own ways and social norms.
Complying (CP)	Students are sufficiently motivated and interested, then actively attending, and committing themselves to the Thainess phenomena. They respond to the phenomena with compliance; that is, they respond to the phenomena consistently with a group or people, social norms, or their teacher's suggestions without their own reasons or full acceptance of the necessity for responding.
Receiving (RC)	Students are sensitized to the existence of Thainess phenomena; that is, they are willing to receive or to attend to them. Students are oriented to learn what teachers intend that they will, and then they bring to each phenomena a point of view which may facilitate or hinder their recognition of Thainess to which the teachers are trying to sensitize them. Students perceived the phenomena without conscious or purposive effort.

Note. The term phenomena refers to situations, stimuli, activities, objects, persons, objects, or state of affairs.

Figure 2.4 *The Thainess construct map*

Level	Description of Student
Expanding (EP)	Students are concerned about the benefits that others will earn from their behaviors about Thai citizenship. Thus, they try to convince others to their cause and try to expand the value of Thai citizenship to others with various forms of thinking, feeling, and behaving.
Valuing (VL)	Students conceptualize values and use their own criterion of worth to assess values in the Thai citizenship phenomena to them. They are satisfied to take the value of Thai citizenship as part of their belief; thus, their behaviors are sufficiently consistent and stable in various situations, from desired to undesired situations, and from simple to complex situations. This means they are holding the value and concerns about the benefits that they will earn from behaving.
Accepting (AC)	Students are concerned about the worth of Thai citizenship and seek to understand how to behave appropriately. They are willing to gain more understanding about the Thai citizenship phenomena. Students accept the worth of Thai citizenship as part of their beliefs and behave based on their own ways and social norms.
Complying (CP)	Students are sufficiently motivated and interested, then actively attending, and committing themselves to the Thai citizenship phenomena. They respond to the phenomena with compliance; that is, they respond to the phenomena consistently with a group or people, social norms, or their teacher's suggestions without their own reasons or full acceptance of the necessity for responding.
Receiving (RC)	Students are sensitized to the existence of Thai citizenship phenomena; that is, they are willing to receive or to attend to them. Students are oriented to learn what teachers intend that they will, and then they bring to each phenomena a point of view which may facilitate or hinder their recognition of Thai citizenship to which the teachers are trying to sensitize them. Students perceived the phenomena without conscious or purposive effort.

Note. The term phenomena refers to situations, stimuli, activities, objects, persons, objects, or state of affairs.

Figure 2.5 *The Thai Citizenship construct map*

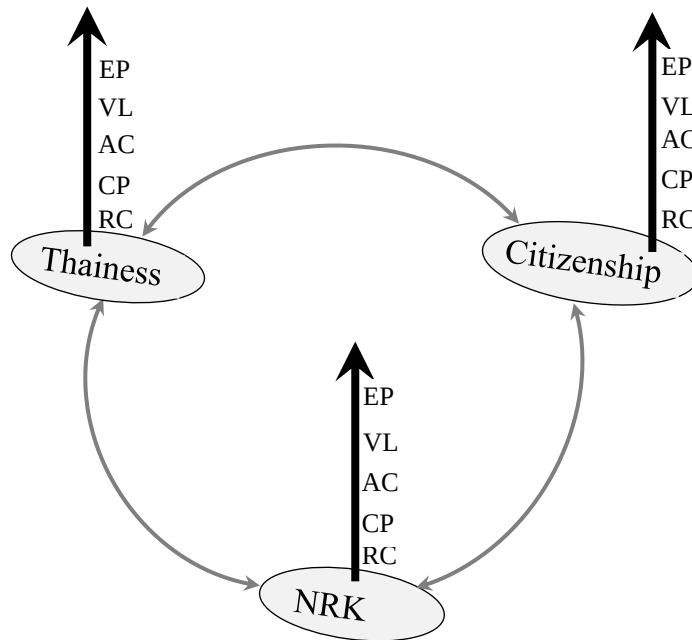


Figure 2.6 *The graphical presentation of the definition of Thai community attachment*

2.2 Developing Situational Items and Scoring Guide

After developing the three construct maps of TCA, the next step is designing items. Results from the literature review show that social desirability such as derived from social norms, culture and tradition, is influential of people's attachment to community (Bar-Tal & Staub, 1997; Kelman, 1997; Johnson, 1997; Primoratz, 2007, 2017). People from different societies and cultures perceive and express their attachment to community differently. For instance, in Western countries, community attachment gives more emphasis on individualization but less on social identification; while, in Asian and African countries, community attachment gives more emphasis on social identification but less on individualization (Reykowski, 1997). Moreover, community attachment is usually learned from varied situations, especially in educational settings where students learn to think, feel, and behave about community attachment from situations at inside and outside schools through inculcation from teachers and parents (Ben-Amos, 1997; Feshbach & Sakano, 1997). These findings support the idea that the TCA should be measured by situational items. .

There were 90 situational items in constructed-response format created to measure the TCA in each dimension (30 items each). All situations used in the items were derived from a focus group of 15 middle school teachers about everyday events and activities to make sure that the situations were suitable for grade 9th students. Each item was designed to measure only one dimension, but it can measure more than one level of the community attachment attribute, depending on the nature of the situation used in that item. All items asked about how a student thinks and feels and how they decide to behave in each situation. Some example items (translated from Thai language) are shown in Figure 2.7.

NRK dimension:

During the weekend, Jane goes shopping with her mother at a farmer's market in her community. She walks behind her mother. Now it is 8 o'clock, she hears the national anthem. Some people do not stop for the anthem including her mother. If you are Jane,

(A) what do you feel about people who do not stop shopping and stand to respect the national anthem? Please describe.

.....

.....

.....

(B) what would you like to do with the people? Please describe.

.....

.....

Thainess dimension:

David is an exchange student from the U.S., and he is Pavit's classmate. Other classmates usually enjoy teaching David to know Thai bad words. They laugh and have fun when David speaks those words, but Pavit never teaches any bad words to him. If you are Pavit,

(A) what do you think and feel about other classmates' behaviors? Please describe.

.....

.....

.....

(B) what would you like to do when your classmates try to teach the bad words to David? Please describe.

.....

.....

Thai Citizenship dimension:

Kittiya is watching news on television. The reporter interviews a foreign traveler at the airport in Bangkok whether he will visit Thailand again. The traveler says, "Sure, I will come back here again. Thailand is a wonderful country. I love Thai food and Thai culture. But many places I visited were dirty and sometimes I saw Thai people did not follow traffic rules". If you are Kittaya,

(A) what do you think and feel about the traveler's answer? Please describe.

.....

.....

.....

(B) what would you like to do after knowing the traveler's feedback? Please describe.

.....

.....

Figure 2.7 Examples of constructed-response TCA item

Constructed-response version of the items were piloted with 283 students (about 100 students per 30 items without time limit conditions during their homeroom hours, then the actual responses to the items were analyzed to obtain response patterns of thinking-feeling-doing combinations for each item to create options for the item. At this stage, the 90 constructed-response items were transformed to 70 selected-response items whose options originated from actual student responses. Moreover, the options of each item were classified into the levels in either the NRK, Thainess, or Thai Citizenship construct maps. It was not necessary for each item to have the same number of options and not necessary to have options to measure all five levels in the construct map. Some options within the same item can measure the same level of possessing the community attachment. This procedure shows that the situations and options define the TCA attribute, according to the operational perspective, in details, not just the operational definition and the scope of the content domains. The classification of the options which are developed from the empirical data and classified into the theoretical structure of student's community attachment establishes the relationship between the theoretical rational structure and the empirical rational structure, according to the representationalism.

Before the 70 items were used in the actual administration, five teachers were interviewed to gain information for improving the 70 items and ten students were chosen to participate in think-aloud interviews to make sure that the students understand the items clearly and use the intended attributes to respond to the items. There were 20 items on the NRK domain, 25 items on the Thainess domain, and 25 items on the Thai Citizenship domain. Some example items in selected-response format are shown in Figures 2.8 to 2.10 and the item specification of the 70 items is shown in Figure 2.11.

During the weekend, Jane goes shopping with her mother at a farmer market in her community. She walks behind her mother. Now it is 8 o'clock, she hears the national anthem. Some people do not stop for the anthem including her mother. If you are Jane, what do you think and feel about the people who do not stop shopping and stand to respect the national anthem and what do you want to do?

A. It is difficult to tell people to stop and respect the anthem. However, at least, I can do it as an example for people.

B. I will tell my mom to stop shopping. We will respect the anthem together, but I am afraid to tell others.

C. Even if there is no one respecting the anthem, I will be the only one to do.

D. It is a normal situation where people do not respect the anthem. Sometimes I do not want to respect the anthem outside my school.

E. There are several ways to demonstrate my love of the nation. If I do not respect the anthem, I think no one will put me in jail.

F. People at the market are shopping. They might not notice that. If most people respect the anthem, I will do so.

Level	Short description of student	Response to the item	Score
Expanding (EP)	Students are concerned about the benefits that others will earn from their behaviors about the NRK.	B	4
Valuing (VL)	Students are holding the value and concerns about the benefits of NRK that they will earn from behaving.	A, C	3
Accepting (AC)	Students accept the worth of NRK as part of their beliefs and behave based on their own ways and social norms.	na.	2
Complying (CP)	Students respond to the NRK phenomena consistently with a group or people, social norms, or their teacher's suggestions without their own reasons.	F	1
Receiving (RC)	Students bring to the NRK phenomena a point of view which may facilitate or hinder their recognition of the NRK to which the teachers are trying to sensitize them.	D, E	0

Figure 2.8 *An examples of selected-response NRK item and response scores*

David is an exchange student from the U.S., and he is a classmate with Pavit. Other classmates usually enjoy teaching Thai bad words to David. They laugh and have fun when David speaks those words, but Pavit never teaches any bad words to him. If you are Pavit, what do you think and feel about other classmates' behaviors and what do you want to do?

A. It is usual for David to speak bad words to his close friends. But I would like him to understand that those words cannot be used to others, and I would like to emphasize that Thai language is beautiful; thus, he should learn other good words as well.

B. It is the nature of adolescents who tend to be so playful. Although those bad words should not be taught, every language has some bad words. Thus, David needs to know.

C. Classmates should not teach the bad words to David. They should teach him to use the language correctly and appropriately in order to understand the identity of Thai language.

D. David is not a Thai. If he uses the language inappropriately, there is no one to blame him; and classmates often speak bad words to their close friends also. Thus, David needs to know some bad words in order to understand what friends are talking about.

E. I have to warn classmates not to teach bad words and convince them to use a proper way to teach Thai language so that David will use Thai language properly, not to be a wrong model for other foreigners.

Level	Short description of student	Response to the item	Score
Expanding (EP)	Students are concerned about the benefits that others will earn from their behaviors about the Thainess.	A, E	4
Valuing (VL)	Students are holding the value and concerns about the benefits of Thainess that they will earn from behaving.	C	3
Accepting (AC)	Students accept the worth of Thainess as part of their beliefs and behave based on their own ways and social norms.	B	2
Complying (CP)	Students respond to the Thainess phenomena consistently with a group or people, social norms, or their teacher's suggestions without their own reasons.	na.	1
Receiving (RC)	Students bring to the Thainess phenomena a point of view which may facilitate or hinder their recognition of the Thainess to which the teachers are trying to sensitize them.	D	0

Figure 2.9 An examples of selected-response Thainess item and response scores

Kittiya is watching news on television. The reporter interviews a foreign traveler at the airport in Bangkok whether he will visit Thailand again. The traveler says that “Sure, I will come back here again. Thailand is a wonderful country. I love Thai food and Thai culture. But many places I visited were dirty”. If you are Kittaya, what do you think and feel about the traveler’s answer and what would you like to do?

A. The image of the country is important. Thai people should keep the country clean. At least, people should keep their neighborhood clean first.

B. I cannot do anything because the country has not been cleaned for a long time. The traveler needs to accept this situation.

C. I and other people have to pick up trash in public space. We should not think that only local governments should do that. This problem can be solved easily if we work together.

D. It is embarrassing when foreigners tell us that our country is dirty. If I have free time, I will join a voluntary activity to clean public space.

E. I agree that some places are not clean. But there are many nice places that travelers should choose to visit. I cannot clean every place. If I have time, I will do it.

Level	Short description of student	Response to the item	Score
Expanding (EP)	Students are concerned about the benefits that others will earn from their behaviors about the Thai citizenship.	C	4
Valuing (VL)	Students are holding the value and concerns about the benefits of Thai citizenship that they will earn from behaving.	A	3
Accepting (AC)	Students accept the worth of Thai citizenship as part of their beliefs and behave based on their own ways and social norms.	D, E	2
Complying (CP)	Students respond to the Thai citizenship phenomena consistently with a group or people, social norms, or their teacher’s suggestions without their own reasons.	na.	1
Receiving (RC)	Students bring to the Thai citizenship phenomena a point of view which may facilitate or hinder their recognition of the Thai citizenship to which the teachers are trying to sensitize them.	B	0

Figure 2.10 An examples of selected-response Thai Citizenship item and response scores

EP (4)	D	D	E	E			C	B	E	E		D	DE	BE	A
VL (3)	C E	E	B		DE	CD E		DE	AB	C	BE	BE	A		C
AC (2)		C	C	BD	B	A	BE				D	C		D	BD
CP (1)	A	A	A	A	C		A	C			C	A	C	C	
RC (0)	B	B	D	C	A	B	D	A	CD	AB D	A		B	A	E
Item no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Label	N1	N2	N3	T1	T2	T3	C1	C2	C3	T4	N4	N5	N6	N7	N8

EP (4)		AC DE	AE	D	A		E	E	D			DE		E	E
VL (3)	BE		C	AB	BC E	CD	AB C	CD	AB C	CD E	DE	C	DE	D	D
AC (2)	AC	B	B	C	D	AE		A			B		C		C
CP (1)				E							A	D	B	AB C	AB
RC (0)	D		D			B	D	B	E	AB	C	B	A		
Item no.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Label	T5	N9	T6	N10	N11	C4	C5	C6	N12	C7	T7	T8	T9	T10	C8

EP (4)	DE	D		B		D	C	C	E	E		C	E	C	E
VL (3)	C	CD	DE	D	E	B	E	DE	D	D	D	D	CD	DE	D
AC (2)			C	E	D	CE	B	B	B	C		A			
CP (1)	A	B	B	AC	C	A	A	A	A	B	B		B	B	
RC (0)	B	A	A		AB		D		C	A	AC	B	A	A	AB C
Item no.	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
Label	C9	C10	C11	T11	C12	T12	T13	C13	T14	T15	N13	N14	N15	N16	C14

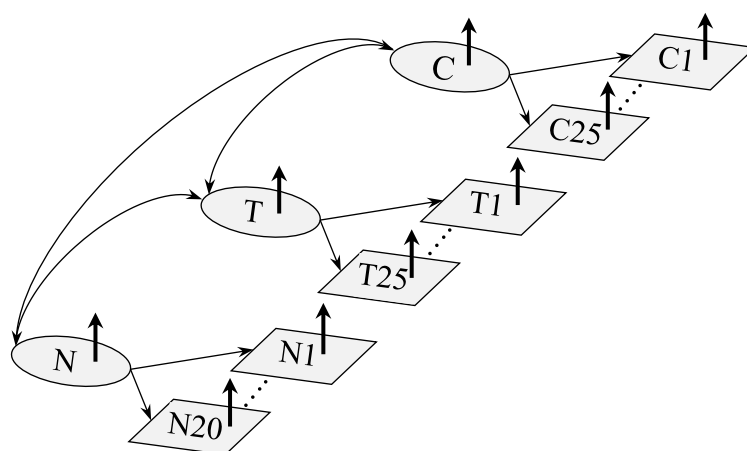
EP (4)		E	D		EF	E	D	D	E		E	E	F	E	
VL (3)	DC	BD	A	AC D	BC D	D	BC E	C	BD	CE D	AD	CD	ED	B	CD E
AC (2)						BC		E	A		C			CD	
CP (1)	E	A						A		A	B	A	AB		
RC (0)	AB	C	BC	B	A	A	A	B	C	B		B	C	A	AB
Item no.	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Label	N17	C15	C16	C17	C18	T16	C19	C20	T17	T18	C21	T19	T20	T21	T22

EP (4)	E	D	D		AD		E	D	C	D
VL (3)	CD	B	CB	BD	B	BC E	D	BE	D	B
AC (2)	B	C			E	A	C	A		C
CP (1)		A		C			AB		B	
RC (0)	A		A	A	C	D		C	A	A
Item no.	61	62	63	64	65	66	67	68	69	70
Label	T23	T24	C22	C23	T25	C24	C25	N18	N19	N20

Note. Numbers 1 to 4 shown in parentheses indicate scores being assigned to student's responses. N refers to the Nation, Religion, and King dimension; T refers to the Thainess dimension; and C refers to the Thai Citizenship dimension.

Figure 2.11 *The TCA items specification*

Figures 2.8 to 2.10 show example items in selected-response format of the three dimensions as well as the relationship between the options and the levels of the construct maps. The levels of the construct maps and the options of the items relate to the specific mathematical structure, according to the representational perspective. Different levels of the TCA and options can be assigned numbers of 0, 1, 2, 3, and 4 to them. The options which meet the lowest Receiving level should be scored as 0. When the options represent the higher levels, their scores should be 1 for the Complying, 2 for the Accepting, 3 for the Valuing, and 4 for the Expanding levels, respectively. According to Figures 2.8 to 2.10, Figure 2.12 shows that the construct maps are situated with respect to the items to establish the relationship between the levels of the TCA and responses to the items and the relationship between the content of attributes and the content of item options.



Note. N, T, and C refer to the nation, religion, and the king (NRK), Thainess, and Thai Citizenship dimensions, respectively.

Figure 2.12 *The graphical presentation of the construct maps situated in dimensions and items*

At this stage, the theoretical connections between the levels of construct maps, the description of student attributes, the items that measure the intended dimensions, and the options that measure the intended TCA levels have been established. Figure 2.13 shows an example of the causality-and-inference relationship between the Thainess dimension and responses to two of the Thainess items. In this figure that is modified from Wilson (2005), the causality goes from the Thainess dimension (as indicated by an oval) to the Thainess items (as indicated by squares). It is assumed that students A and B have different degrees of the Thainess attribute. Student A has a higher degree, at the Expanding level, while Student B has a lower one, at the Receiving level. The different levels of the attribute are causes of the responses to the items T1 to T25. The degree at Expanding level causes Student A to choose options E of item T1 and option A and D of item T25 that these three options were created to measure the Expanding level; while being at the Receiving level causes Student B to choose option C of item T1 and option C of item T25

which were created to measure the Receiving level. When two students respond to the items, their responses imply the underlying Thainess attribute and the different levels of the attribute. Thus, the direction of inference is back from the Thainess items to the Thainess dimension. When applying this explanation to the other dimensions, it is important to notice that the TCA attributes are well-defined according to the operational and representational perspectives.

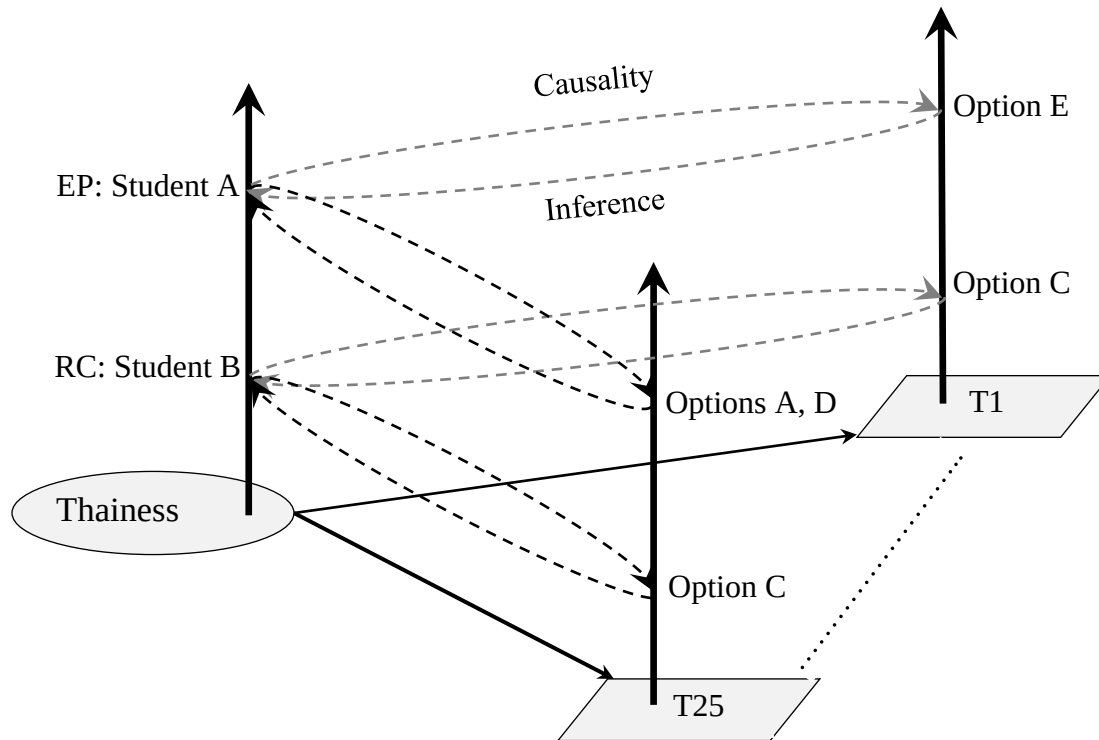


Figure 2.13 *The graphical presentation the relationship between degree of attribute and item responses*

2.3 Establishing Measurement Models for Parameter Estimation

The last step of designing the TCA measurement using the construct modeling approach is choosing measurement models, that is item response theory (IRT) models. An IRT model describes the relationship between the person attribute and the probability of a certain response on an item (Embretson & Reise, 2000; Hambleton et al., 1991; Wilson, 2005). At this step, two measurement models were developed for estimating the psychometric properties of the items and the students. The first model is a unidimensional item response theory (UIRT) model, as shown on panel (a) of Figure 2.14, where all 70 items measure a theoretically non-separate dimension of the TCA and the second is a multidimensional item response theory (MIRT) model, as shown panel (b) of Figure 2.14, where three sets of the items measure different dimensions as specified

previously. This multidimensional model is called a between-item multidimensional model (Wang et al., 1997).

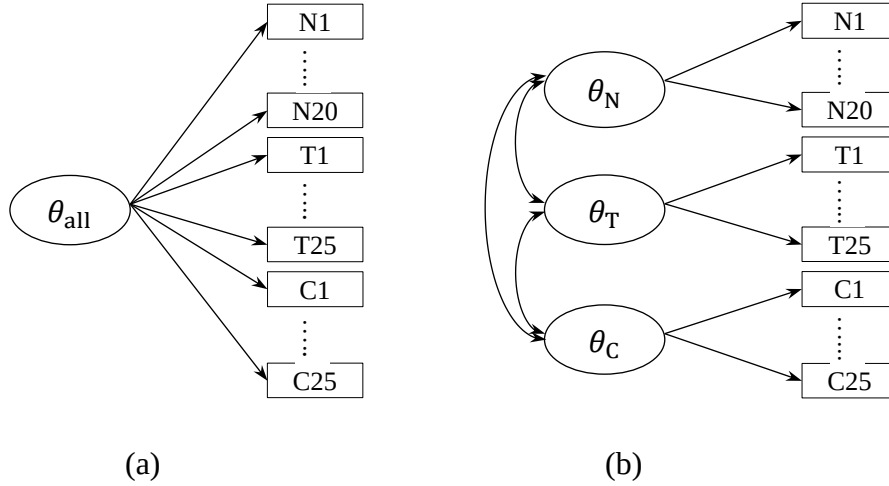


Figure 2.14 *Measurement models*

Under the psychometric approach in modeling item responses multidimensionally based on Rasch measurement known as the Multidimensional Random Coefficients Multinomial Logit (MRCML) model framework (Adams et al., 1997), a between-item MIRT model involves estimating student's attributes when an instrument or test consists of several dimensions and each item measures a particular dimension. That is unidimensionality exists at the item level, while multidimensionality exists at the test level.

In a multidimensional instrument, students use multiple attributes to respond to items. Their responses are the most important ingredient for a MIRT model to demonstrate the relationship between the probability of a response and the location of a person on a multidimensional attribute matrix. The location of person n ($n = 1, 2, \dots, N$) is represented by the row vector $\theta_n = (\theta_{1n}, \theta_{2n}, \dots, \theta_{dn}, \dots, \theta_{Dn})$, where D is the number of latent dimensions needed to describe the attributes when responding to items. In other words, the elements of the vector provide coordinates on the D dimensions for the location of person n . When person n responds to an item which relates to a between-item MIRT model, person n will draw on the relevant dimensions to determine the response to the item.

The between-item MIRT model is extended from the standard Rasch model (Rasch, 1980). The Rasch model is the simplest model in item response theory and assumes all items in an instrument measure the same latent dimension—known as the unidimensionality assumption (Hambleton et al., 1991; Wang et al., 1997). The IRT model being applied to the response data assumes that there are more than two possible values: the partial credit model (PCM; Masters, 1982). The PCM, as specified in Equation 2.1, models the probability of person n responding in category j of item i as a function of the person attribute θ_n and step parameters δ_{ij} ,

$$P(x_{in} = j | \theta_n, \delta_i) = \frac{\exp \sum_{j=0}^J (\theta_n - \delta_i)}{\sum_{j=0}^{m_i} \exp \sum_{j=0}^J (\theta_n - \delta_i)}, \quad j = 0, 1, \dots, m_i \quad (2.1)$$

where $\sum_{j=0}^0 (\theta_n - \delta_i) \equiv 0$, and where $\theta_n \sim N(0, \psi)$, and m_i is the total number of steps in item i ($m_i + 1$ is the number of categories); so that $\sum_{j=0}^{m_i} \exp \sum_{j=0}^J (\theta_n - \delta_i)$ is the sum of the numerators for all categories.

Equivalently, the PCM can be specified in Equation 2.2 as the log-odds of the probability of person n responding in category j of item i to the probability of responding in category $j - 1$ as a function of the person attribute θ_n and step parameter δ_{ij} ,

$$\log \frac{P(x_i = j | \theta_n, \delta_i)}{P(x_i = j - 1 | \theta_n, \delta_i)} = \theta_n - \delta_{ij}, \quad j = 0, 1, \dots, m_i \quad (2.2)$$

where $\theta_n \sim N(0, \psi)$, and m_i is the total number of steps in item i .

These expressions show the simple relationship between the person location and the item location in the Rasch unidimensional model. For Rasch multidimensional modeling, the MRCML model is a direct extension of the unidimensional random coefficient multinomial logit model (Wilson & Adams, 1995; Wilson & Wang, 1995) and is built up from basic conceptual building-blocks. Within the MRCML model, it assumes a set of d ($d = 1, 2, \dots, D$) latent dimensions that determine person attributes underlying the person responses. Then, the log-odds of the probability of a person n response in category j of item i (P_{ij}) compared to category $j - 1$ (P_{ij-1}) as a linear function of latent attribute on dimension d (θ_d) in Equation 2.3, and the relative difficulty of item i in category j (δ_{ij}) can be modeled in a between-item model as (Briggs & Wilson, 2003):

$$\log \left(\frac{P_{nij}}{P_{ni(j-1)}} \right) = \theta_{nd} - \delta_{ij}. \quad (2.3)$$

The θ_{nd} in Equation 2.3 indicates that the person latent attribute is measured on the attribute dimension d mapped onto item i . The mapping, then, would indicate that item i is related to only dimension d , and the value of δ_{ij} would indicate whether it is relatively easier or harder for a person to be classified as achieving category $j - 1$ compared to category j .

In conclusion, the MRCML model can be implemented and provides a convenient and concrete method for estimating and locating person attributes and the item step parameters on the same scale. This means person attributes can be interpreted based on the item step parameters or their associated values, such as the item thresholds.

The selected-response items were tested with the intended 980 students. Their response data was divided into two datasets—a calibration dataset for estimating psychometric parameters and another validation dataset for evaluating the parameters and results. Psychometric parameters were estimated using ConQuest 2.0 software (Wu et al., 2007). Validity evidence according to the Standards for Educational and Psychological Testing (AERA et al., 2014) were

collected to support the plausibility of interpreting the measurement results and the appropriateness of use of the measurement results.

The response data to the TCA items contain certain categories for null responses. These null categories occur by the measurement design when each item does not need to have options to measure the TCA levels from the lowest level, Receiving, to the adjacent higher level, Complying, to the next level, Accepting, and so forth. For instance, item N1 has no option to measure the Accepting level, but its adjacent levels, Complying and Valuing, and item N9 has options that measure only Accepting and Expanding. In order to deal with this measurement situation, the downcode technique (Wilson & Masters, 1993) is used by reducing the score of each category above the null category by one. Thus, the original response data with null categories responses was transformed to the downcoded data before the parameter estimation. For example, for item N1, option C and E will be reduced their score from 3 to 2, and option D from 4 to 3. For item N9, option B will be reduced its score from 2 to 0, and options A, C, D and E will be reduced its score from 4 to 1.

(1) Model Fit Comparison

The unidimensional model is viewed as a nested model in the between-item multidimensional model which contains additional covariance associated with the random effects of the latent dimensions. In order to compare the relative fit of the nested models, a likelihood ratio (LR) test is used. The LR statistic is equal to the difference in deviance (G^2) between two models. This test ensures whether a given model fits significantly better or worse than a competing model. It has been shown that the null distribution of the LR statistic is approximately a χ^2 distribution with degrees of freedom that equals the difference in the number of parameters estimated by the two models. To compute the LR test, the differences of the G^2 values (G^2_{diff}) and the degree of freedom (df_{diff}) of the two models are taken as Equation 2.4:

$$G^2_{diff} = G^2_s - G^2_l, \text{ and } df_{diff} = df_s - df_l, \quad (2.4)$$

where s denotes the smaller model, less restrictive unidimensional model, with fewer parameters and thus more df , whereas l denotes the larger model, more restrictive multidimensional model, with more parameters and fewer df . If the G^2_{diff} value is significant, we conclude that the more restrictive model fits the item response data better than the less restrictive model. That means it makes sense to estimate the additional parameters and to prefer the more restrictive model. Conversely, lack of significance of the G^2_{diff} shows that we do not have evidence to reject the unidimensional model in favor of the multidimensional one, so the additional parameters can be eliminated from the less restrictive model (or fixed to zero) and the more restrictive model can be accepted.

The G^2 values have their own interpretation. Since the G^2 equal $-2 \times \log\text{-likelihood}$, which indicates the amount of remained variation in the response data unexplained by the model (Wilson & Moore, 2012), a low value of G^2 indicates better fit—and that can be interpreted that the model with additional parameters fits better (or provides less unexplained variation in the data) than the model without additional parameters.

Similarly to a χ^2 test applied to a single model, the LR tests applied to the nested models have essentially the same strengths and weaknesses. They are directly affected by sample size, and for a large sample may become significant for even very small difference values (Schermerle-Engel et al., 2003; Werner & Schermerle-Engel, 2010). Thus, Cohen's effect size measure w (Cohen, 1988) is used to quantifying the magnitude of difference in chi-square as Equation 2.5:

$$w = \sqrt{\frac{\Delta\chi^2}{(n \cdot \Delta df)}} \quad (2.5)$$

where $\Delta\chi^2$ and Δdf are the differences in chi-square and df, respectively, of the two models, and n is the number of samples. The values can be interpreted in terms of Cohen's suggested standards for a small as .1, medium as .3 and large as .5 effect

The second set of indexes used in fitting the models is the Akaike information criterion (AIC: Akaike, 1974). AIC offers a relative measure of the information lost when a given model is used to describe reality. It describes the tradeoff between bias and variance in model construction. Given a dataset, several competing models are compared by calculating the likelihoods of their solutions. The greater the likelihood, or lower $-2 \times \log$ -likelihood (G^2) the better the model fits the data (Janssen & de Boeck, 1999). The competing models may be ranked according to their AIC, in this study, and the one having the lowest AIC is the best. The general form of calculating AIC is:

$$AIC = -2 \times (\log - \text{likelihood}) + 2p, \quad (2.6)$$

where p is the total number of estimated parameters in the model.

AIC is not asymptotically consistent as sample size is not used in its calculation. The next index also used to assess the models is Bayesian information criterion (BIC: Schwarz, 1978):

$$BIC = -2 \times (\log - \text{likelihood}) + p(\log n), \quad (2.7)$$

where n is the sample size. BIC penalizes relatively more for the number of parameters and hence favors models with fewer parameters compared to AIC. Models which have small AIC and BIC are selected.

(2) Item Quality Evaluation

In order to investigate the quality of the items, the test statistic of weighted fit (infit) with mean square are used rather than the unweighted fit (outfit) mean square since the infit statistic is more robust than the outfit to unusual observations in terms of their values and numbers (Wright, 1984) and infit is an information-weighted form of the outfit (Wright & Masters, 1990). Under most circumstances, however, infit and outfit values will be similar (Adams & Khoo, 1998).

The mean square is an effect size corresponding to the t-value and associated p-value. High or low (especially high) MNSQ and t-values may indicate differences between observed and expected item characteristic curve (ICC) other than just item discrimination. Thus, the mean

square accounts for the relationship between t-value and item discrimination influenced by sample size in three conditions (Wilson et al., 2006). If the mean square has a value of 1, the observed variance equals the expected variance in the deviation of item discrimination—and that can be interpreted that the slope of the empirical ICC is probably captured well by the slope of the ICC generated by the estimated item parameters. Values greater than 1 indicate an unexpectedly high variance in the deviation of item discrimination—implying that the empirical ICC may be flatter than the estimated ICC, and values less than 1 indicate an unexpectedly low variance in the deviation of item discrimination—implying that the empirical ICC may be steeper than the estimated ICC.

Since there is no absolute criterion for measuring an effect size in the item parameter estimation and specifying the magnitude of item misfitting, a range of weighted mean square values between .75 ($=3/4$) and 1.33 ($4/3$) is widely considered acceptable (Adams & Khoo, 1998). If the fit statistics of an item falls outside the range, the item is considered to be misfitting.

The 95% confidence interval around 1.0 of the infit statistics is also under consideration. If the infit statistic of an item falls outside the confidence interval, the null hypothesis that the response data conforms to the model at the .05 significance level can be rejected. That means the item behaves differently from what was expected based on the estimated item parameters at the .05 significance level. Note that the t-value is sensitive to a large sample size, thus in this study both mean square and confidence interval criteria are used to evaluate the item fit.

(3) Model Invariance

Cross-validity is an empirical property which indicates the degree of model invariance obtained from the data itself across independent subsamples. It gives one more confidence in replicating the overall estimation results (Brown, 2000; Hair et al., 1998; Mellenbergh, 2011) as well as providing evidence to avoid overfitting the model (Kevin, 1991). In this study, it refers to the degree that estimated parameters of the best fit model obtained from the first independent sample do not differ significantly from the ones obtained from the second independent sample. Therefore, the procedure mainly involves the use of the first subsample, called the calibration sample, to produce estimation results, and then to validate the first results by the second subsample, called the validation sample (Hair et al., 1998; Mellenbergh, 2011).

In this study, cross-validation is performed in four simple steps. The first step is to obtain the calibration and validation samples drawn as a structured random sample from the response data. The second step is to use the calibration sample to estimate the psychometric parameters and the variance-covariance matrix. The third step is similar to the second but uses the validation sample instead. At the end of this step, two sets of structural parameters and of the variance-covariance matrix will be obtained from the first and second estimations, respectively. The final step is comparing the estimation results focusing on item step parameters. In order to interpret the differences, the logic of interpreting the effect size of item parameter difference across two samples in a differential item functioning analysis (Paek, 2002) is applied. That means that if the difference in item step parameter between the calibration and validation samples is less than .426, it is not considered large; if this is true for (most or) all items, it is reasonable to conclude that the best fit model shows model invariance a cross-validity properties.

(4) Dimensional Alignment Analysis

One constraint required for model identification of the MRCML model framework is that the abilities for each dimension are centered on zero (Briggs & Wilson, 2003; Feuerstahler & Wilson, 2019). Because of this constraint, no direct comparison between dimensions can be made (Schwartz, 2012). In fact, the item specification and the description of student's attribute are different in terms of the number of items measuring each dimension and the different complexity of TCA. The mean of the item difficulties for each dimension, thus, might not have a common mean of zero.

Delta dimensional alignment (DDA) is a technique for comparing estimated attributes across dimensions when the best fit model is a multidimensional model (Schwartz, 2012; Feuerstahler & Wilson, 2019). In order to examine how the levels of TCA are associated across dimensions, the DDA technique is used to transform the item step parameters obtained after running an initial multidimensional analysis. The item difficulties (δ) of different dimensions are transformed by using the means ($\text{mn}(\hat{\delta}_{ud})$) and standard deviations ($\text{sd}(\hat{\delta}_{ud})$) of the subsets of the items for each dimension ($d = 1, 2, \dots, D$), which are calculated from the unidimensional model. This technique relies on an assumption that multiple dimensions are positively correlated to some degree and the metric of one single dimension is reasonable to use to align the metrics across all the multiple dimensions (Schwartz, 2012; Feuerstahler & Wilson, 2019). The three steps of the DDA technique are described below.

The first step is to run the unidimensional analysis to obtain the item location estimates and then compute the mean ($\text{mn}(\hat{\delta}_{ud})$) and standard deviation ($\text{sd}(\hat{\delta}_{ud})$) for each subset of items by dimension according to the item specifications. It makes sense to believe that the unidimensional model in this step is a theoretical model underlying the overall TCA attribute. The second step is to run a multidimensional analysis to obtain another set of item location estimates, then compute the standard deviation ($\text{sd}(\hat{\delta}_{\mathcal{M}d})$) for each subset of items by dimension. Note that the mean ($\text{mn}(\hat{\delta}_{\mathcal{M}d})$) for each dimension is set to zero in the estimation, due to the model identification constraint. Using the item difficulties obtained from both analyses in Equations 2.8 and 2.9,

$$\hat{r}_d = \frac{\text{sd}(\hat{\delta}_{ud})}{\text{sd}(\hat{\delta}_{\mathcal{M}d})}, \quad (2.8)$$

$$\hat{s}_d = \text{mn}(\hat{\delta}_{ud}) - \hat{r}_d \text{mn}(\hat{\delta}_{\mathcal{M}d}) \quad (2.9)$$

then transform the multidimensional item estimates using Equation 2.10 for item locations and Equation 2.11 for step parameters τ :

$$\tilde{\delta}_{i(d)} = \hat{r}_d \delta_{id} + \hat{s}_d \quad (2.10)$$

$$\tilde{\tau}_{ij(d)} = \tau_{\mathcal{M}ijd} \left(\frac{\text{sd}(\hat{\delta}_{ud})}{\text{sd}(\hat{\delta}_{\mathcal{M}d})} \right) \quad (2.11)$$

The final step is to run another multidimensional analysis using these transformed item and step parameters as anchored values, in order to obtain new values of the student attribute. Since the parameters are anchored, the multidimensional model can estimate the student attribute without the previous constraints on the mean of student attribute parameters. As a result, the student attributes can be compared across the dimensions.

Chapter 3

Results and Validity Evidence for the Thai Community Attachment Assessment

3.0 Introduction

The purpose of this chapter is to present the measurement results of the assessment of Thai community attachment in relation to the measurement philosophies of operationalism, representationalism, and pragmatism, and the sources of validity evidence based on response processes, test content, internal structure, testing consequences, and relations to other variables. However, the last type of validity evidence on the list is not available this time to compare with other results relevant to the TCA assessment. It is possible that this evidence will be made in the future measurement.

3.1 Response Processes Validity Evidence

Since the validation using response processes was conducted during item development, I will show its results in this section and the other three types of validity evidence will be presented throughout the chapter. The validity evidence based on response processes was obtained from ten students who participated in think-aloud interviews to make sure that the students understood the items clearly and used the intended attributes to respond to the items. The results suggested ways to eliminate potential ambiguities and revise the situations of items. For example, there were needs to provide clearer definitions of “attachment”, not “nationalism”, and “feeling”, not “emotion.” The results were also used to revise options and adjust the descriptions of students in the construct maps. Thus, this validation helped improve the content of the attributes corresponding to the operational perspective.

3.2 Sample

In the spring of 2019, a total number of 980 grade 9 students from large urban schools in three provinces in Thailand was assigned to take the TCA items within at most three hours during their homeroom periods. The TCA items in the paper-based format were split into three parts according to its three dimensions; that is there were 20 items for the NRK part, 25 items for the Thainess part, and 25 items for the Thai Citizenship part. The students can complete each part within one hour of their homeroom period each day. There were no omitted or blank responses in the data. This indicates that the students had very little limitation of time. The sample consisted of 542 (55.31%) female students and 438 (44.69%) male students. There were 840 (85.71%) students who lived in the same communities for at least 10 years. Moreover, there were 767 (78.27%) students who identified themselves as knowing their communities very well.

The sample was divided randomly into two subsamples. The first sample of 490 students was used to calibrate item and student parameters and evaluate both unidimensional and between-item multidimensional models. Another sample of 490 students was used to evaluate the invariance of the parameters and revise the measurement results. These two subsamples suffice to yield more precise and robust estimates with small standard errors and a greater power fit analysis than a smaller sample (Bond & Fox, 2007; Linacre, 1994; Wright & Tennant, 1996).

The results are divided into 10 sections according to the investigation procedure as follows.

3.3 Model Fit Comparison

The result in Table 3.1 suggests that the multidimensional model fit significantly better than the unidimensional model ($df = 5$, $chi-squared = 31.05$, $p < .01$), which provides evidence that the subsets of the TCA items measure three distinct attributes of the NRK, Thainess, and Thai Citizenship. Even though the model fit of the three-dimensional model does not show a large effect size or high practical significance, the results can be used as validity evidence based on internal structure of the TCA to support the scope of TCA content domains according to operationalism.

Table 3.1 Model fit statistics comparison

Model	No of parameters	Deviance	AIC	BIC
Unidimensional	190	42053.07	42433.07	42564.21
Multidimensional	195	42022.02	42412.02	42546.61

The next investigation of the differences between the models at the interpretation level is correlation analysis of student attributes. A correlation matrix in Figure 3.2 has two sets of scatter plots and coefficients. The first set above the diagonal was obtained from the consecutive unidimensional models, while the second below the diagonal is obtained from the multidimensional model.

All correlation coefficients are positive. The multidimensional model produced much higher correlation coefficients (between .92 and .95) than the ones from the consecutive models (between .54 and .62). This result informs us that the consecutive models are not appropriate for estimating student's TCA attributes. This can also be hypothesized that there will be inconsistency in student attribute estimates when treating the NRK, Thainess, and Thai Citizenship dimensions independently. The positive correlation coefficients from the multidimensional model indicate that students who have high estimated attributes on one dimension will have high values on second dimension.

Actually, the multidimensional model producing such high correlations might be an argument for accepting the unidimensional model, but these high correlations suggest that there shared cognitive processes or a common attribute among the three dimensions while students tried to understand the situations and options in different contexts and domains, and/or they use the "attachment" as a common attribute to respond to the TCA items. However, the variations in

the three dimensions has not been completely explained by the same common attribute or the same cognitive processes. Thus, for this study, the multidimensional model is still useful for the future investigations. This is similar to the 2015 PISA (an international large-scale assessment) that the correlations of student performance scores between reading, mathematics, and science are high (between .75 and .88) as expected (OECD, 2017) but there was still some separation and variation between each of the dimensions unexplained.

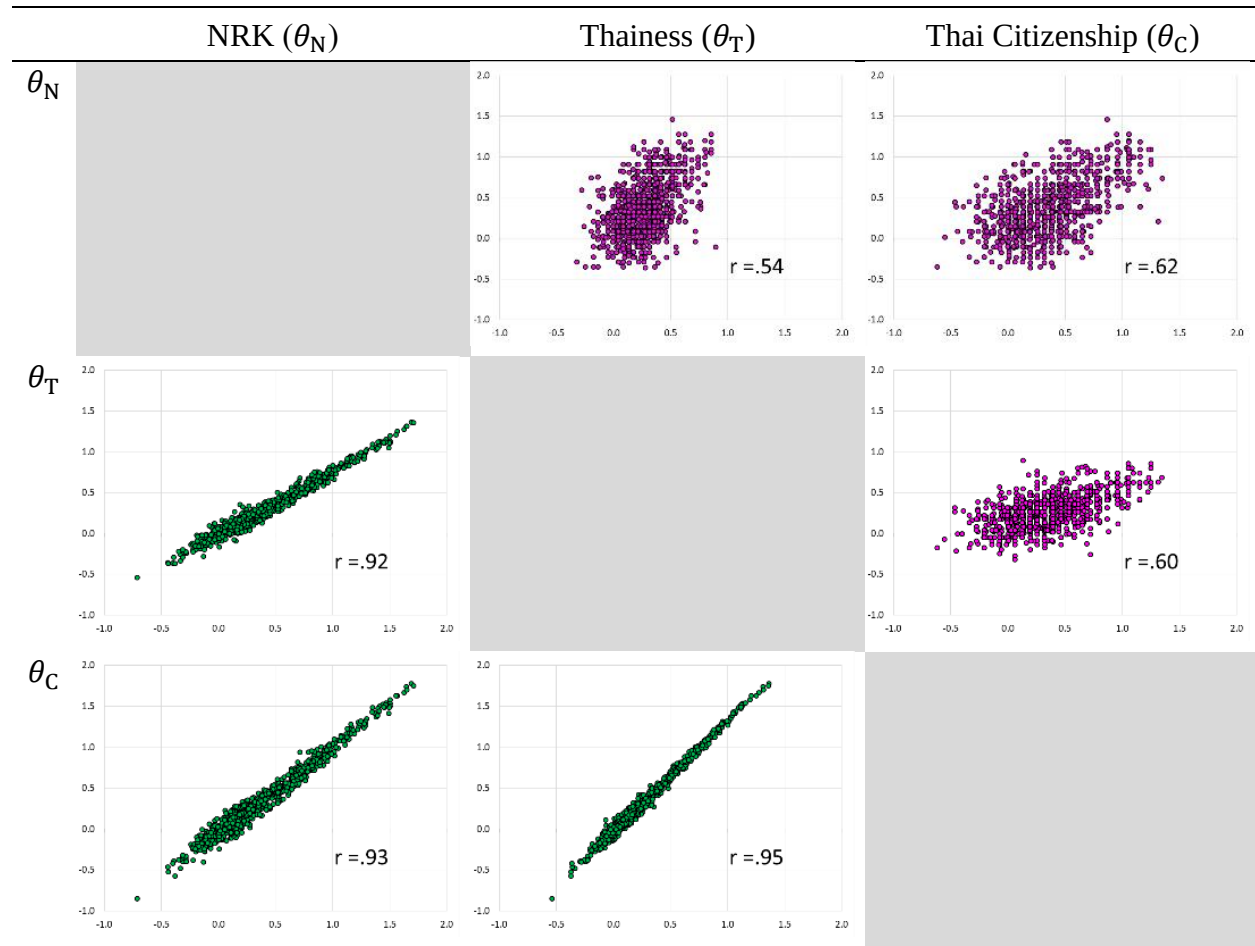


Figure 3.1 *Correlation matrix of student's TCA attributes*

3.4 Reliability Comparison

According to the indication of model fit, it makes sense to expect that there will be differences between the models at the interpretation level; thus, several consequential investigations were established to show evidence to support the plausibility of the TCA score interpretation. The next investigation is a posteriori over person variance based on plausible values (EAP/PV) reliability comparison. The EAP/PV reliability is explained variance according to an estimated model divided by total person variance. A small EAP/PV reliability indicates that the explained variance from the model is small while the total person variance is large; thus, the student attribute estimates are inconsistent. On the other hand, a large EAP/PV reliability indicates that the explained variance from the model is large while the total person variance is small; thus, the student attribute estimates are consistent. Considering the reliabilities of the better fit model in Table 3.2, they are obviously higher than the consecutive unidimensional models (Briggs & Wilson, 2003) that treat and model the sum of observed scores on items associated with their relevant NRK, Thainess, and Thai Citizenship dimensions independently as three unidimensional models. The reliabilities from the multidimensional model meet the acceptable rule of internal consistency reliability of .80 (Nunnally, 1978; Taber, 2018) that the better fit model reflects 20%, 20% and 18% measurement error of student attribute estimates in the NRK, Thainess, and Thai Citizenship dimensions, respectively. This result supports the idea of the three-dimensional structure of the TCA.

Table 3.2 *EAP/PV reliability comparison*

Dimension	Consecutive unidimensional model	Multidimensional model
NRK	.64	.80
Thainess	.51	.80
Thai Citizenship	.65	.82

3.5 Standard Error of Measurement Analysis

The investigation of standard error of measurement (SEM) of student attribute estimates serves the concept that each student attribute location is an estimate which contains a degree of uncertainty and the uncertainty is usually characterized using the stand error of the student attribute location. The investigation results are the SEM graphs in Figure 3.1. The SEM graph compares each student attribute estimate with its SEM; that means the SEM provides information about how accurate the student attribute estimate is. The interpretation is as follows: It is assumed that the exact student's attribute location is normally distributed around the estimated student's attribute location with standard deviation equal to the SEM (Wilson, 2005). For example, if a location of student's attribute estimate is .40 logits with a SEM of .15, then the student attribute estimate can be interpreted that there is 95% certain that the exact student's attribute location lies in an interval of $.40 \pm 1.96 \times .15$ logits (.11, .69). This confidence interval means that the exact student attribute location might be anywhere between .11 and .69 logits. Thus, the smaller the SEM, the smaller the width of the confidence interval, and then the greater accuracy and reliability in the student attribute estimate.

Figure 3.2 shows a set of SEM graphs obtained from EAP estimation. The top graph with red dots was obtained from the unidimensional model, and the rest of the graphs with blue dots for the NRK dimension, yellow dots for the Thainess dimension, and purple dots for the Thai Citizenship dimension, were obtained from the multidimensional model. Each SEM graph comes with two sets of descriptive statistics about SEM and student's attribute estimate. When considering the four SEM graphs, their relationships are similarly flat line shapes. Low SEMs are found with low student attribute estimates and a little higher SEMs are found with high student attribute estimates. Thus, according to the four graphs, there are no large differences in SEM for students whose estimated attributes are either low or high, which is a good feature, due to the sound plan for item generation.

When considering the descriptive statistics of SEM of the unidimensional and multidimensional models, there are no different stories found. In the multidimensional model, the SEM ranges from .14 to .29, which is considered small values to give high reliable attribute estimates. Therefore, the results suggest that the TCA items can contribute well to the estimation of student attribute locations accurately between -.85 and 1.78 logits. This result corresponds to validity evidence based on internal structure to support the findings about higher reliabilities of the better fit multidimensional model according to the operational perspective

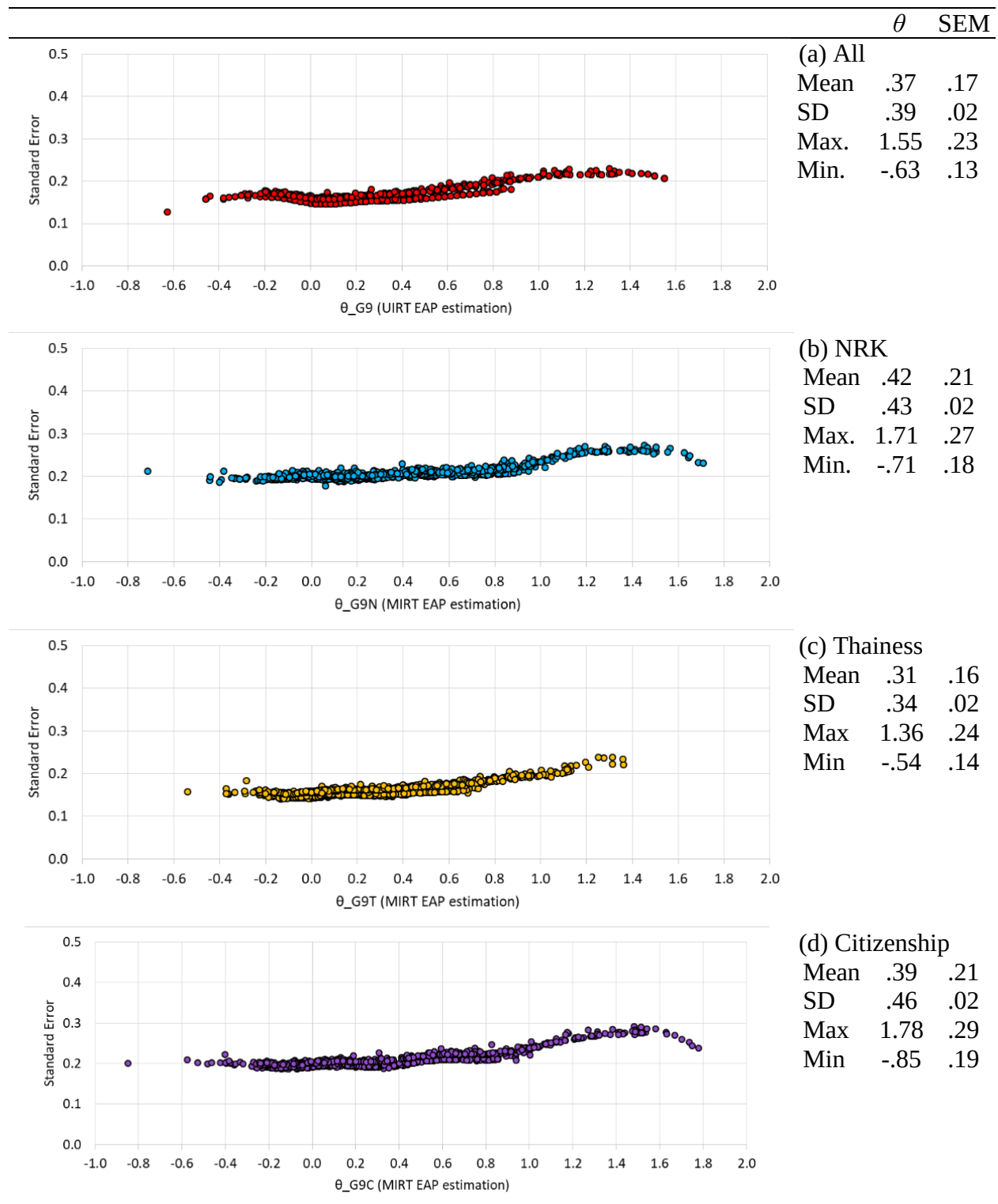


Figure 3.2 *Graphs and descriptive statistics of standard error of measurement*

3.6 Item Statistical Quality Evaluation

The statistical quality of the items was evaluated by infit mean square statistics. Figure 3.3 shows the infit statistics obtained from the better fit multidimensional model for average item parameters on the left panel and for item step parameters on the right panel. Theoretically, a value of 1 of the infit statistics means that the observed squared residual of an item equals the expected square residuals. The infit statistics of most TCA items both average item parameters and step parameters are close to the value of 1 with a range of .90 to 1.10. This result shows that there are small differences between observed variance and expected variances in student's attributes; that means their item characteristic curves generated from observed response data are similar to the ones generated from the student's attribute estimates, then when student's attribute increases the probability to achieve the higher levels of TCA in each dimension also increase. The result also shows that most items discriminate well between students who have low and high levels of the NRK, Thainess, and Thai Citizenship dimensions.

However, the infit statistics are directly affected by a large sample, even small differences may become statistically significant. When evaluating the magnitude of item misfit by considering the effect size in the item parameter estimation, all infit statistics fall inside the .75 to 1.33 acceptable range (Adams & Khoo, 1996) as indicated by the yellow zones, except for item T24. This is a good result and reflects to fact that these items were developed from several qualitative analyses and the item options were developed from actual responses to constructed-response format. The infit statistics of item T24 as 1.46 can interpret that there is 46% more variation in the response data than modelled by the multidimensional model.

As discussed earlier, the response levels, which are represented by the step parameters, were designed to relate directly to the levels of the NRK, Thainess, and Thai Citizenship construct map. For example, item T15 with options that measure all five Thainess levels will have scores of 0, 1, 2, 3, and 4 assigned to responses, and the item will have four item steps. The first step discriminates the Receiving level from the Complying level; the second discriminates the Complying from the Accepting; the third discriminates the Accepting from the Valuing; and the fourth item step discriminates the Valuing level from the Expanding level. Therefore, the individual item steps fit is more interesting than the item average fit. On the other hand, the infit statistics for the average item location gives some indication of the overall quality of an item. A high infit mean square value suggests there are problems with the design of an item warranting investigation. Thus, item T24 with the highest infit mean square of 1.46 outside the acceptable range and thus needs revision or deletion. The evaluation results of item quality above can be used as validity evidence based on internal structure. Almost all items capture the variation in response data well.

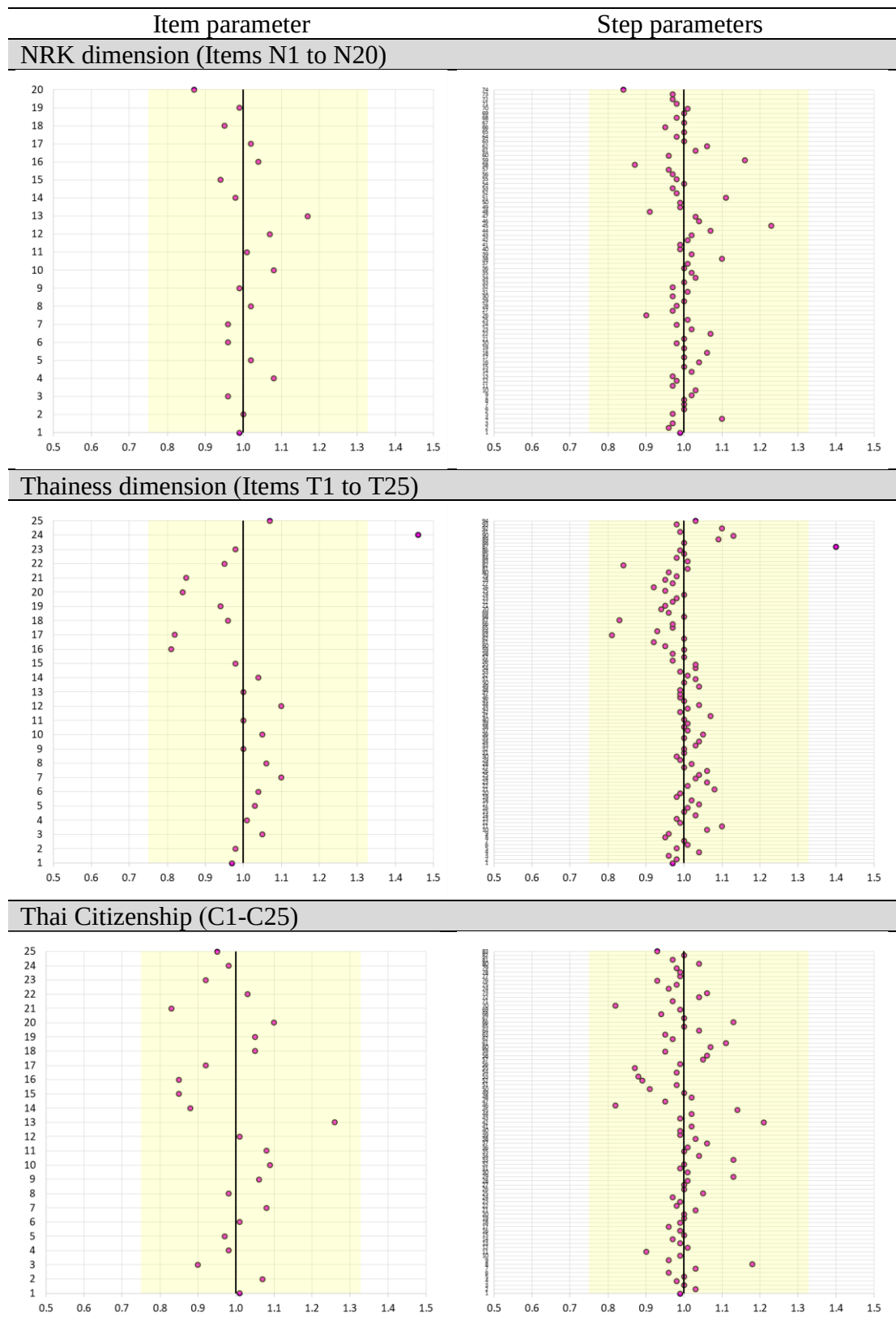


Figure 3.3 *The TCA Infit mean square graph*

3.7 Item Calibration Validation

The next investigation is about how an TCA item can be explained by the TCA item parameters and the student attribute estimates. The Wright map is an especially useful source of information to serve this investigation. The Wright maps as shown in Figures 3.4 to 3.6 were obtained from the multidimensional model using expected a-posteriori (EAP) estimation (Wang et al., 1997), and modified from the original outputs of ConQuest 2.0 (Wu et al., 2007). These Wright maps allow me to relate the empirical data to the theoretical construct maps. The dashed vertical lines express the latent attributes in logits (or the log of odds as shown in Equations 2.2 and 2.3 in chapter 3), which relate the latent attributes to the probability of responses. The logits provide the units of the attribute (as mentioned in earlier this chapter) and are scaled on the left of the dashed vertical lines, under 'Logit'. The dashed vertical line also relates to the developmental continuum structure, running from possessing less TCA attributes, the Receiving level, at the bottom to possessing more TCA attributes, the Expanding level, at the top of the up arrow. Individual students with different degrees of possessing the TCA attributes will be mapped to the logits scale to display their developmental structure.

On the left of the dashed vertical lines, under 'Student', the student attribute locations on the logits scale are indicated by 'X'. All the 'X's form a histogram on the Wright maps, and these show the distribution of the student attribute estimates. The right-hand side of the dashed vertical lines, under 'Item', shows the average item locations by using item numbers to refer to item labels, under 'Item label' on the Wright maps. Item numbers 1 to 20 refer to items N1 to N20 for the NRK dimension in Figure 3.4; item numbers 21 to 45 refer to items T1 to T25 for the Thainess dimension; and item numbers 46 to 70 refers to items C1 to C25 for the Thai Citizenship dimension.

The double dashed vertical lines are used to separate the item locations from the item step locations, and they can be used as the logits scales to indicate the locations of item steps also. The numbers .1 to .4 next to the item numbers refer to the sequence of item steps, under '1st', '2nd', '3rd', and '4th', respectively. The item steps are grouped into four categories using different color boxes: the first, second, third, and fourth groups are represented with yellow, green, blue, and pink, respectively. The groups of item steps relate to the levels of the NRK, Thainess, and Thai Citizenship construct maps. The idea of grouping the item steps is to investigate how well the same item steps that relate directly to the same intended level are located and how the different item steps locations are ordered.

When considering Figures 3.4 to 3.6, the item steps locations are expected to group and step from the bottom left to the top right of the Wright maps to show the order of first item steps locations to the fourth one, corresponding to the order of levels in NRK, Thainess, and Thai Citizenship construct maps. However, there are some inconsistencies found in the item steps locations that do not order as expected by the construct map. For example, in Figure 3.4, the NRK dimension, item step 9.4 was expected to locate closely to the fourth steps of other items. It should not locate at the same levels with item steps 12.1 and 8.2. Similarly, item step 13.1 should not locate at the same level with the center of the fourth item steps. Moreover, item steps 13.3 and 11.3 should not locate far from the third step of other items. Overall, this same story holds for the Thainess and Thai Citizenship dimensions in Figures 3.5 and 3.6 as well.

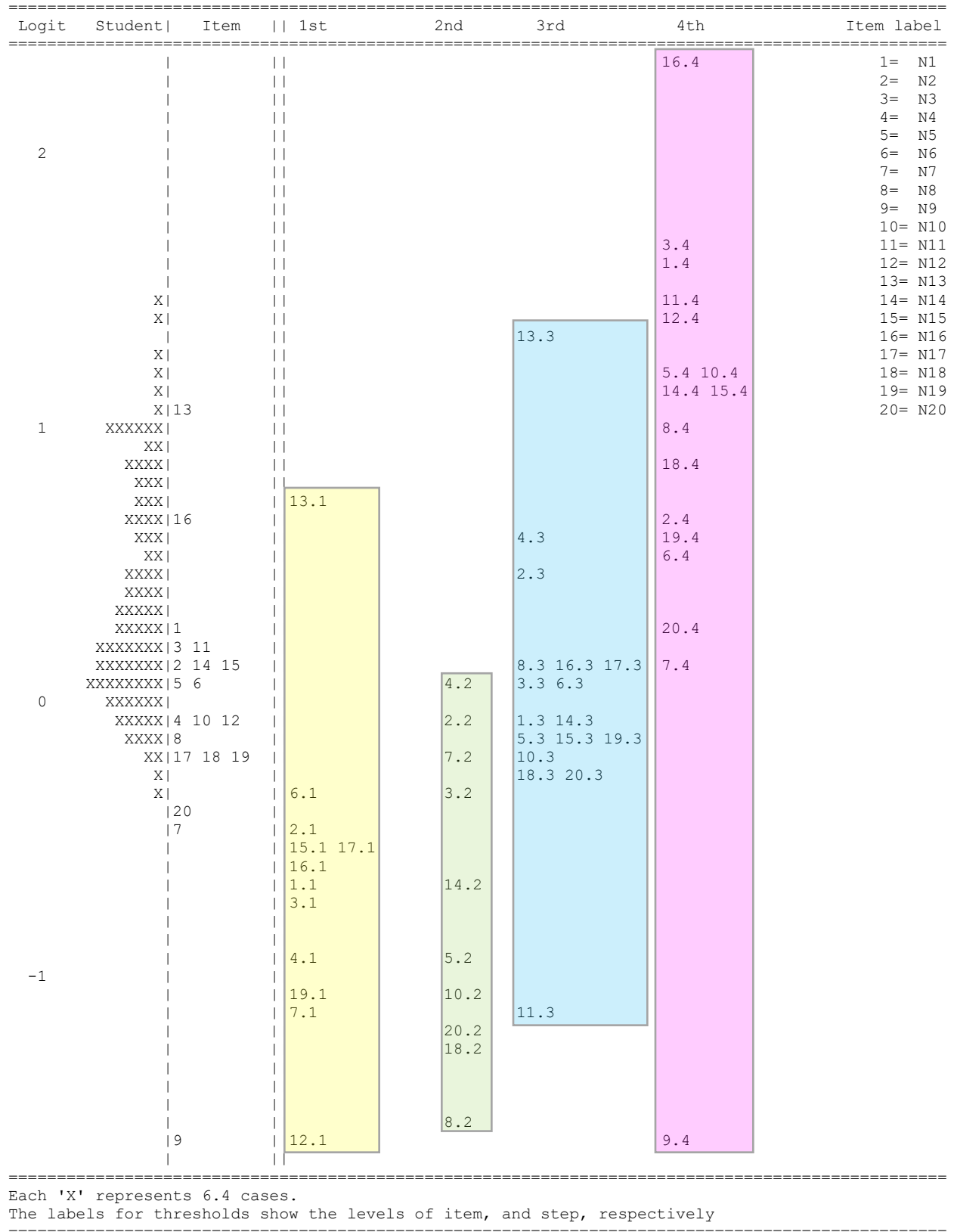


Figure 3.4 The Wright map of NRK dimension

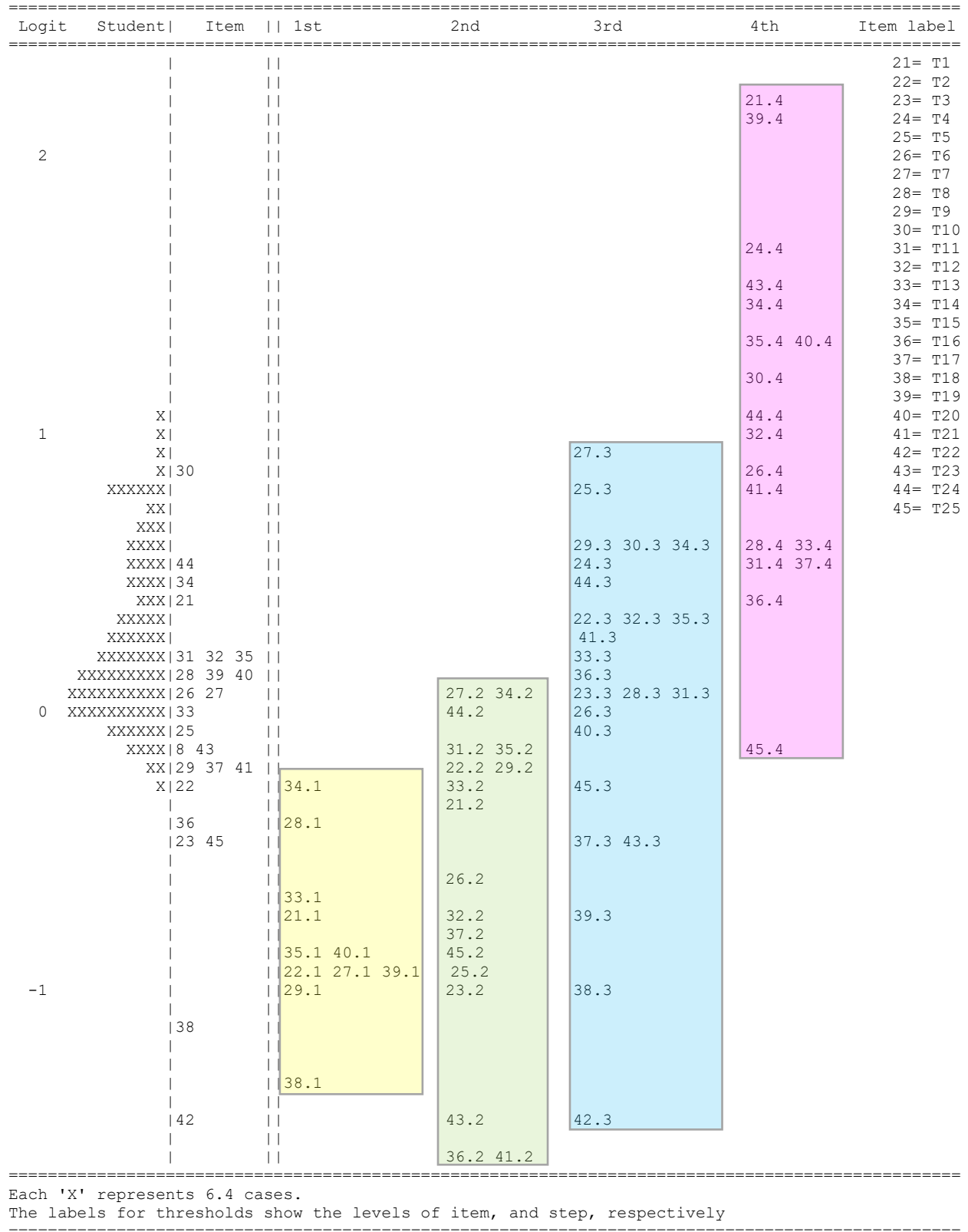


Figure 3.5 The Wright map of Thainess dimension

The Wright maps in Figures 3.4 to 3.6 show that there are three sets of problematic items because their item step locations behave inconsistently from their groups. These items negatively affect the order of levels in NRK, Thainess, and Thai Citizenship dimensions. The first set for the NRK dimension consists of four items including item N13 with step 13.1, item N8 with step 8.2, item N11 with step 11.3, and item N9 with step 9.4. The second set for the Thainess dimension consists of seven items, including item T14 with step 34.1, item T8 with step 28.1, item T23 with step 43.2, item T16 with step 36.2, item T21 with step 41.2, item T22 with step 42.3, and item T25 with step 45.4. The final set for the Thai Citizenship dimension consists of seven items, including item C12 with step 57.1, item C22 with step 67.3, item C5 with step 50.3, item C19 with step 64.3, item C17 with step 62.3, item C18 with step 63.3, and item C16 with step 61.4.

There are two main problematic patterns found in these inconsistent items. The first pattern is that some items ask students to choose between two important situations to do under specific conditions. For example, item N13 asks students to choose between helping their family to run a home business during weekend (lower response scores) and volunteering to a social activity (higher response scores); and item C12 asks students to choose between finishing individual assignments (lower scores) or group assignments (higher scores) when the two deadlines are the same. Students tended to have low response scores on these two items because they mainly considered themselves. This makes the locations of first and second item steps are higher than other in the same item step groups.

The second pattern is that when items use common situations such as respecting traffic regulations (items C17), returning lost items (item C18), and being kind to the poor (T22), students tend to have high response scores. This makes the locations of third and fourth item steps are lower than other in the same item step groups.

In order to improve the TCA assessment, a revision analysis of the better fit multidimensional model was conducted by deleting all 18 hypothesized problematic items as mentioned above. A new set of the revised Wright maps are shown in Figures 3.7 to 3.9. The results show that the locations of same item steps in each dimension are more grouped and the orders of levels in each dimension are clearer. Moreover, there are acceptable overlapping locations between the adjacent levels that reflect the uncertainty of TCA assessment.

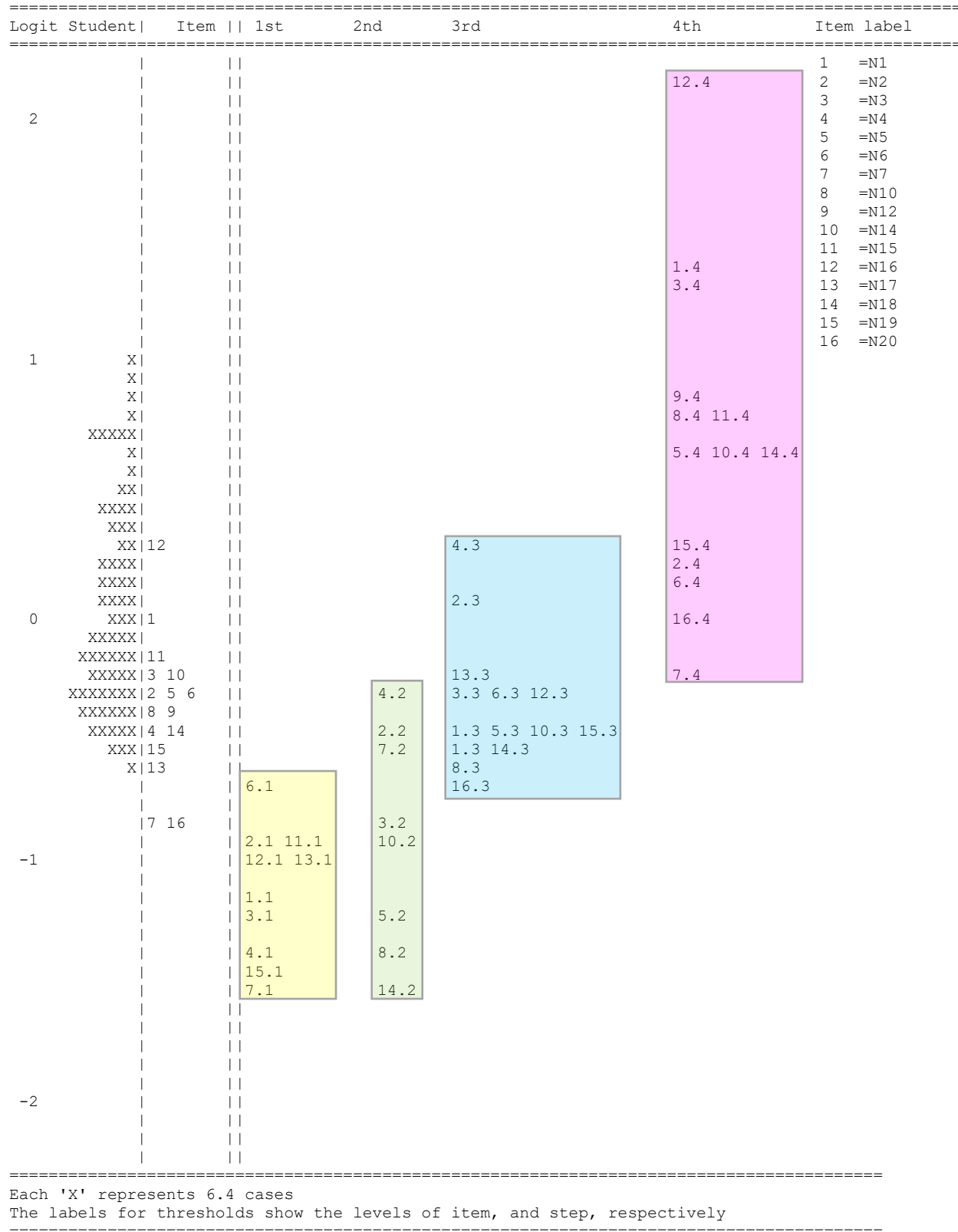


Figure 3.7 The revised Wright map of NRK dimension

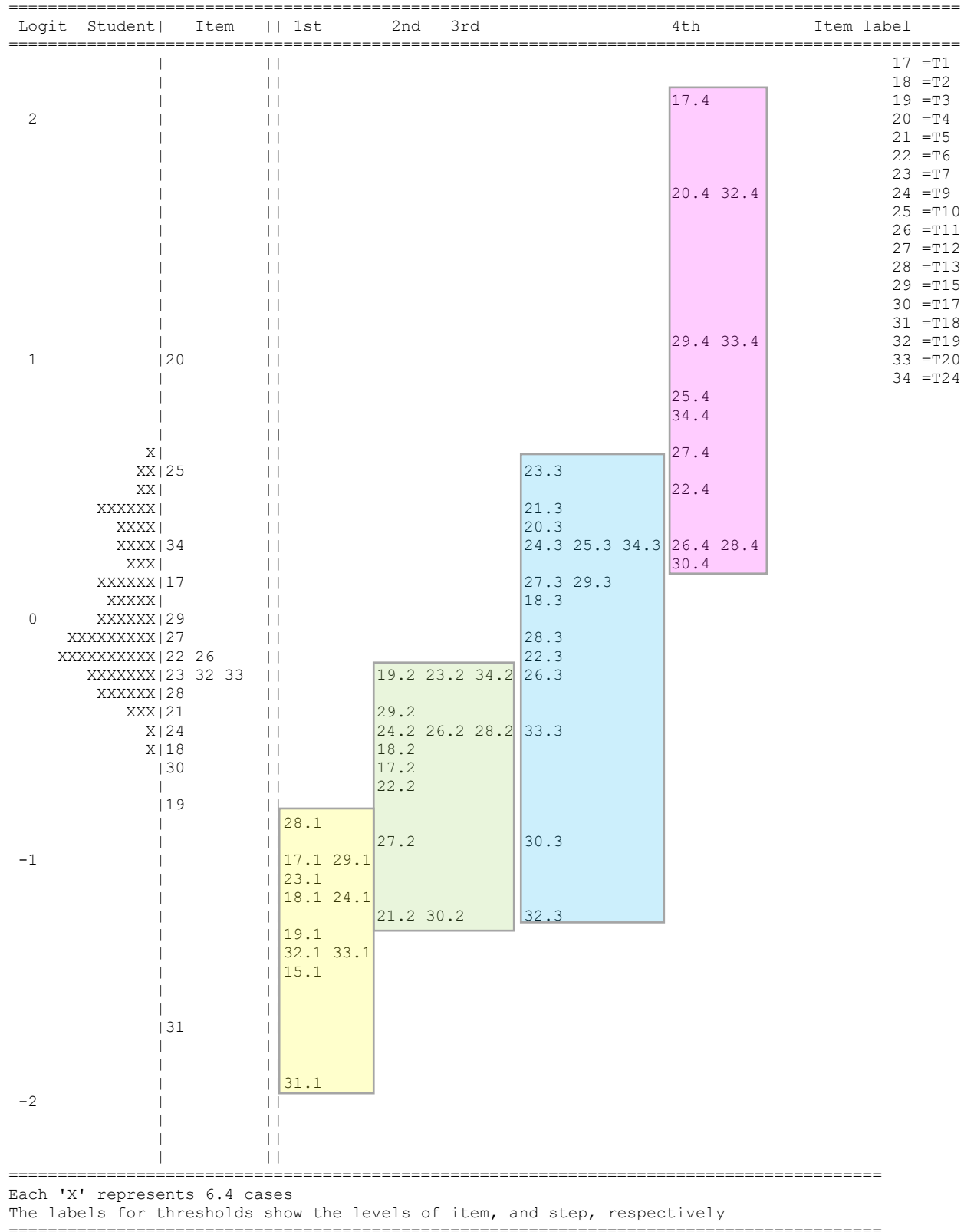


Figure 3.8 The revised Wright map of Thainess dimension

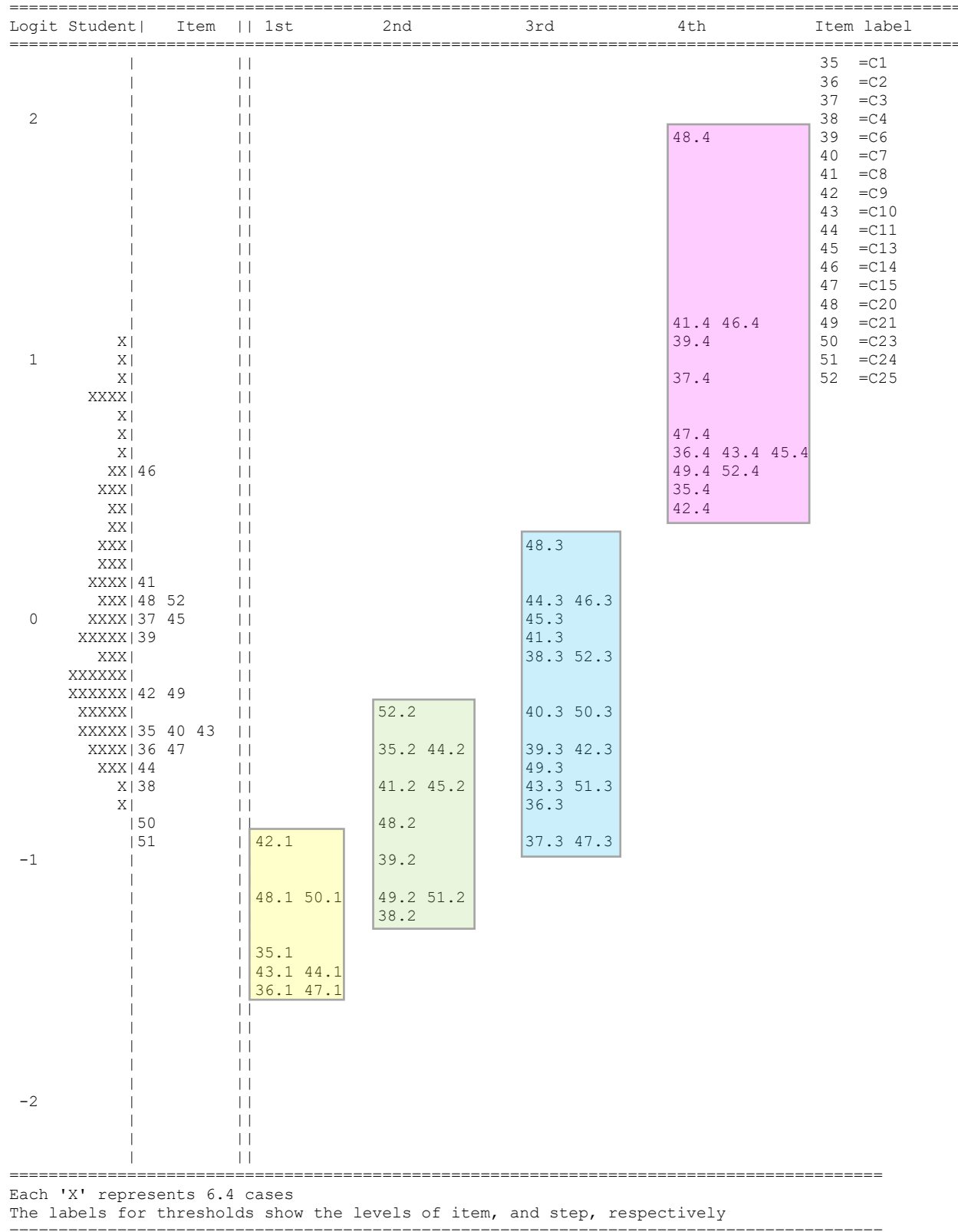


Figure 3.9 The revised Wright map of Thai Citizenship dimension

3.8 Item-Level Validation

Table 3.3 shows the results from investigating each item's consistency with the instrument as a whole by considering the locations of the students' attribute within each score category for a given item. The concept of this investigation is that the mean attribute locations of each category are expected to increase as the score increases. Students with a high TCA level would have a high score on each item. This trend relates directly to the developmental continuum structure situated in the three construct maps and the logits score of the Wright maps. A drop in mean attribute location is highlighted in blue, yellow, and pink for the NRK, Thainess, and Thai Citizenship dimensions, respectively. The many blanks shown in Table 3.3 indicate that the items were designed to not have options to measure that level corresponding to the item specification in the previous chapter.

Table 3.3 *The mean locations trajectories of student attributes in each level of TCA*

Item	Level of NRK					Item	Level of Thainess					Item	Level of Thai Citizenship				
	(score)						(score)						(score)				
	RC	CP	AC	VL	EP		RC	CP	AC	VL	EP		RC	CP	AC	VL	EP
	(0)	(1)	(2)	(3)	(4)		(0)	(1)	(2)	(3)	(4)		(0)	(1)	(2)	(3)	(4)
N1	-.32	-.27		.14	.17	T1	-.20	-.20	.07		.11	C1	-.38	-.17	-.08		.33
N2	-.38	-.26	-.09	.06	.29	T2	-.24	-.13	-.08	.15		C2	-.48	-.31		.08	.16
N3	-.38	-.35	-.24	.13	.28	T3	-.25		.03	.06		C3	-.32			-.03	.40
N4	-.34	-.20	-.20	-.11		T4	-.15			-.05	-.15	C4	-.30		-.28	-.13	
N5		-.33	-.30	-.12	-.10	T5	-.18		-.10	-.11		C6	-.41		-.31	-.14	-.04
N6	-.34	-.26		-.26	.00	T6	-.16		-.20	-.09	-.04	C7	-.25			-.17	
N7	-.39	-.29	-.21		-.09	T7	-.16	-.08	-.19	-.06		C8		-.36	-.30	-.07	-.02
N10		-.30	-.27	-.13	-.13	T9	-.23	-.20	-.09	-.06		C9	-.42	-.32		-.06	-.16
N12				-.19	-.19	T10			-.15	-.09	-.04	C10	-.34	-.32		-.17	-.13
N14	-.37		.01	.35	.39	T11		-.20	-.15	-.10	-.04	C11	-.35	-.28	-.23	-.10	
N15	-.24	-.07		.21	.49	T12	-.09		-.16	-.13	-.06	C13		-.24	-.35	-.13	-.08
N16	-.26	-.05		.46	.07	T13	-.25	-.19	-.18	-.01	.05	C14	-.14			.49	.59
N17	-.15	.21		.30		T15	-.22	-.17	-.17	-.04	.06	C15	-.26	-.46		.15	.65
N18	-.37		-.06	.24	.44	T17	-.31		-.13	.00	.34	C20	-.45	-.41	.17	.46	.50
N19	-.23	-.01		.19	.40	T18	na.	-.26				C21		-.45	-.23	.17	.66
N20			na.	-.04	.46	T19	-.23	na.		.16	.25	C23	-.47	.06		.43	
						T20	-.26	-.13		.20	.39	C24	-.35		-.03	.37	
						T24		.13	.10	.04	.28	C25		-.17	-.17	.18	.66

Note. na. refers that a mean location is not reported if there are less than 10 cases in that score category.

There are 8 (15%) out of 52 items showing the unexpected trends of mean student attribute locations that fail to increase from the Receiving to Expanding levels. The results reveal some problems in the measurement design and procedure of TCA, especially the relationship between the theoretical structure, as shown in the construct maps, and the empirical structures, as

shown in the Wright maps and mean attribute location trajectories of TCA levels. These problems have not received enough attention in most psychological measurements because psychological researchers and psychometricians have relied on only the operational perspective, not the representation and pragmatic ones, when creating and developing their measurements. This justification does not mean that the problems found in the TCA measurement design can be accepted and leave them with no solutions and does not mean that most typical psychological measurements are inaccurate. On the other hand, this justification reflects that evidence for the validity of many psychological measurements show an incomplete picture about their measurement when introducing different perspectives of measurement philosophy to measurement design to gain the new understanding about psychological measurements, the plausibility of interpretation of measurement results, and the appropriateness of use of measurement results.

A strategy can be used to address the problems is deleting some hypothesized items with unexpected item steps parameters. This strategy has been done in the previous section. The items that show unexpected trends of mean attribute locations in Table 3.3 are accepted as the uncertainty in the TCA assessment and the items will not be deleted.

3.9 Model Invariance Investigation and Measurement Revision

There are four steps of the revision using the validation sample. The first step is to delete the items that has been done in the calibration sample in the validation sample also. The second step is to run the multidimensional model using the validation sample with anchoring the variance-covariance matrix and item step parameters obtained from the validation sample. The final step is to compare the estimation results focusing on item step parameters. The difference can be interpreted in terms of the effect size of item step parameter difference in a differential item functioning analysis.

This investigation about model invariance aims to gain more confidence in replicating the overall estimation results, especially item step parameters and student attribute estimates, and to avoid overfitting the multidimensional model. In order to initially examine the invariance of the multidimensional model, a second estimation was conducted using the validation sample without anchoring the variance-covariance matrix obtained from the calibration sample. The results in Table 3.4 show that the values of variance and covariance from both samples are slightly different ranging from .012 to .079.

Table 3.4 *Comparison of variance-covariance matrices of the multidimensional model*

Dimension	NRK	Thainess	Thai Citizenship
Calibration sample			
NRK	.237		
Thainess	.138	.123	
Thai citizenship	.249	.161	.311
Validation sample			
NRK	.307		
Thainess	.156	.111	
Thai citizenship	.231	.130	.232

Table 3.5 *EAP/PV reliability (total sample)*

Dimension	Number of items	Total sample
NRK	16	.77
Thainess	18	.67
Thai citizenship	18	.76

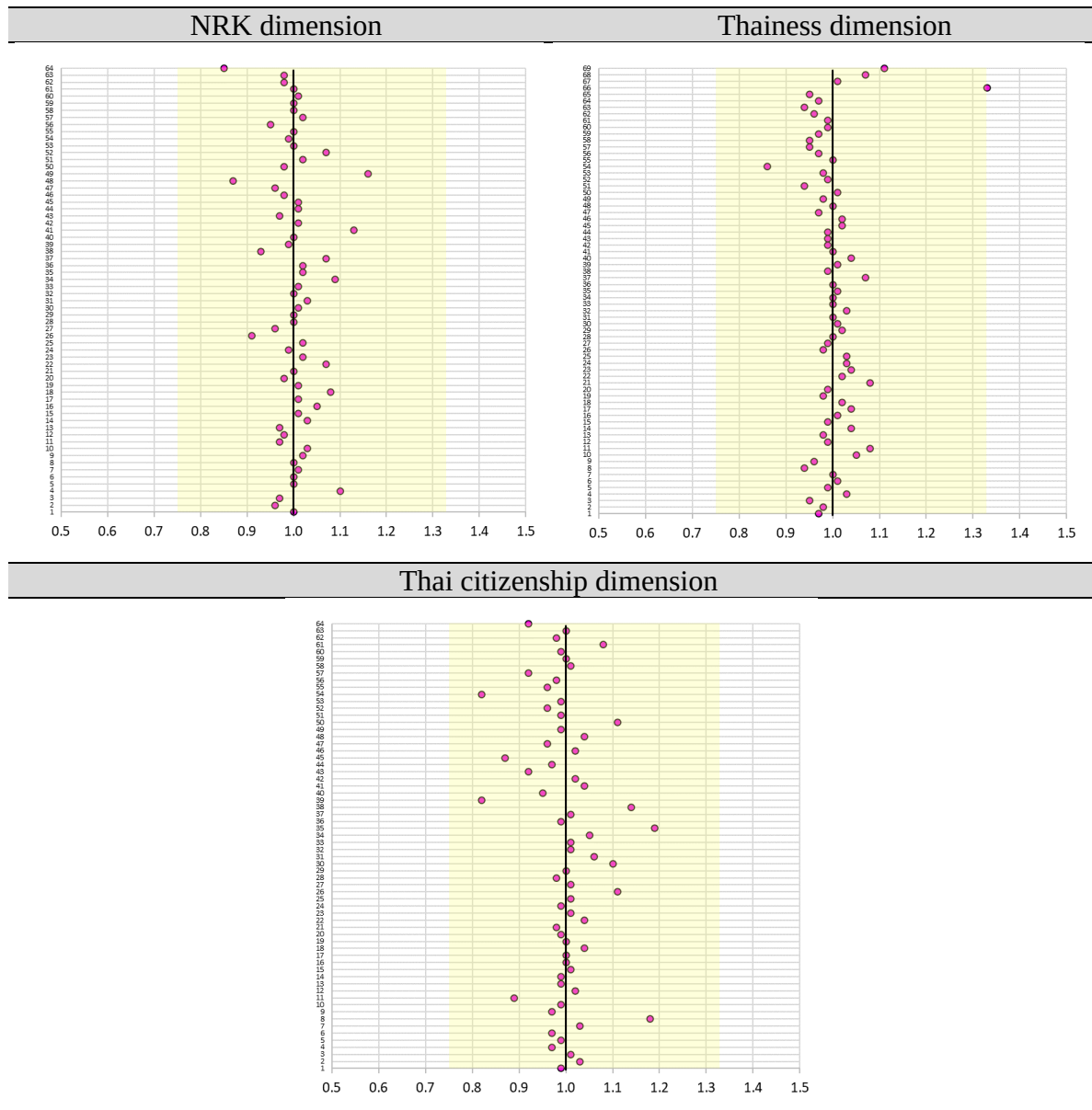


Figure 3.10 *The TCA Infit mean square graph (total sample)*

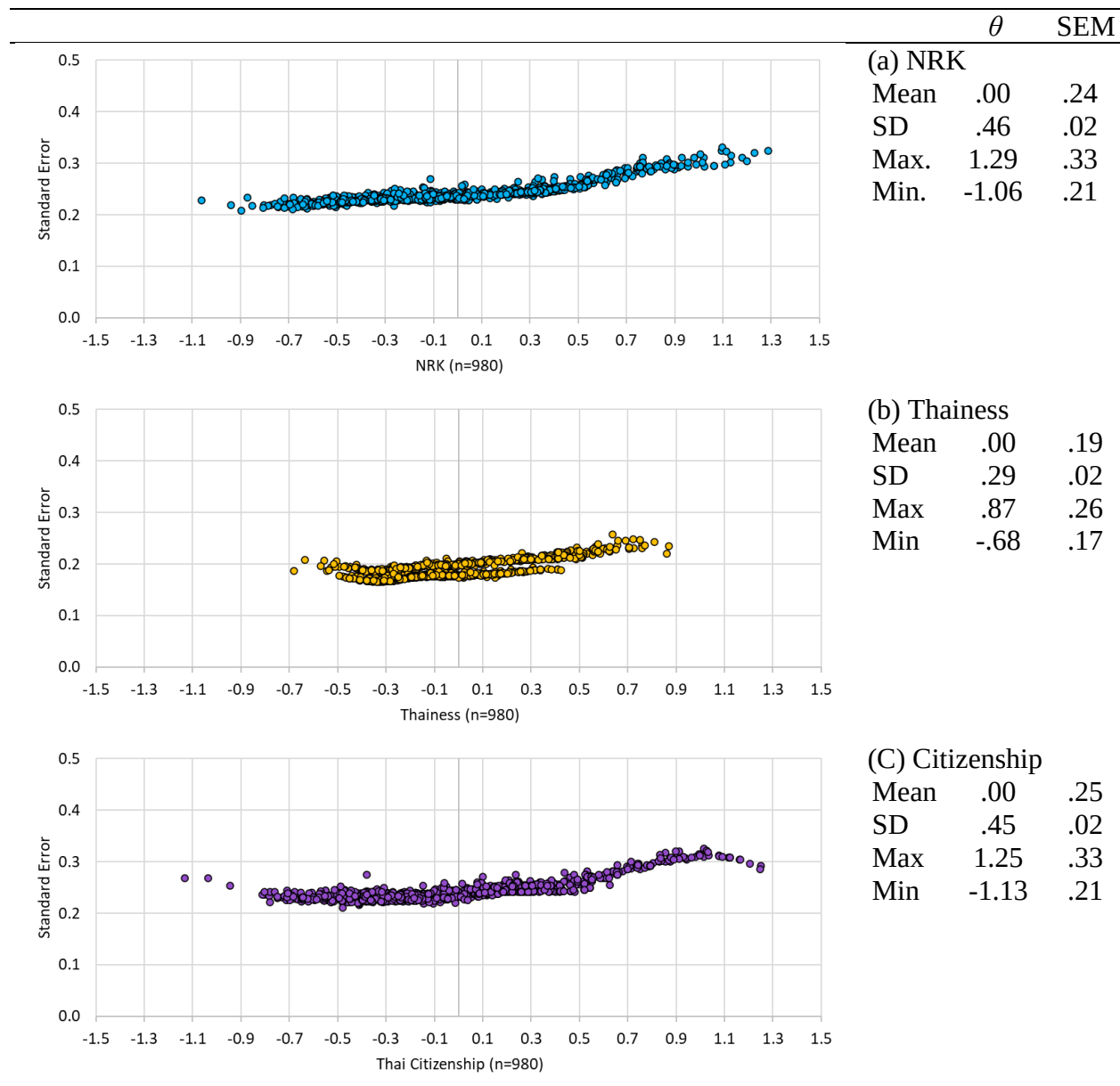


Figure 3.11 *Graphs and descriptive statistics of standard error of measurement (total sample)*

When anchoring the validation sample with the variance-covariance matrix obtained from the calibration sample, the differences in item step parameters are very small, ranging from .00 to .09. They all meet the criteria of insignificant difference, less than .426. Thus, the multidimensional model is quite invariant with respect to item step parameters.

After examining the model invariance using the second sample, the calibration sample and validation sample were combined to get the total sample of 980 students, then used the total sample to estimate the student attribute parameters and item step parameters from the better fit multidimensional model. The estimation results in Figure 3.10 show the infit mean square,

ranging from .81 to 1.33, all lie within the .75 to 1.33 range, and in Table 3.5 show that the EAP/PV reliabilities are slightly lower than the acceptable rule of .80 as result of the item deletion. These results can be used as validity evidence based on internal structure to support the scope and components of TCA according to the operational perspective.

The SEM graphs and the descriptive statistics of SEM in Figure 3.11 show that when using the total sample, the student attribute estimates are accurate and reliable with low SEMs. This result corresponds to the previous analysis and confirms that the remaining TCA items after excluding the unexpected items can contribute well to the estimation of student attributes between -1.20 to 1.30 logits. The SEM graphs and the descriptive statistics of SEM can be used as validity evidence to support the accuracy of multidimensional model according to the operationalism.

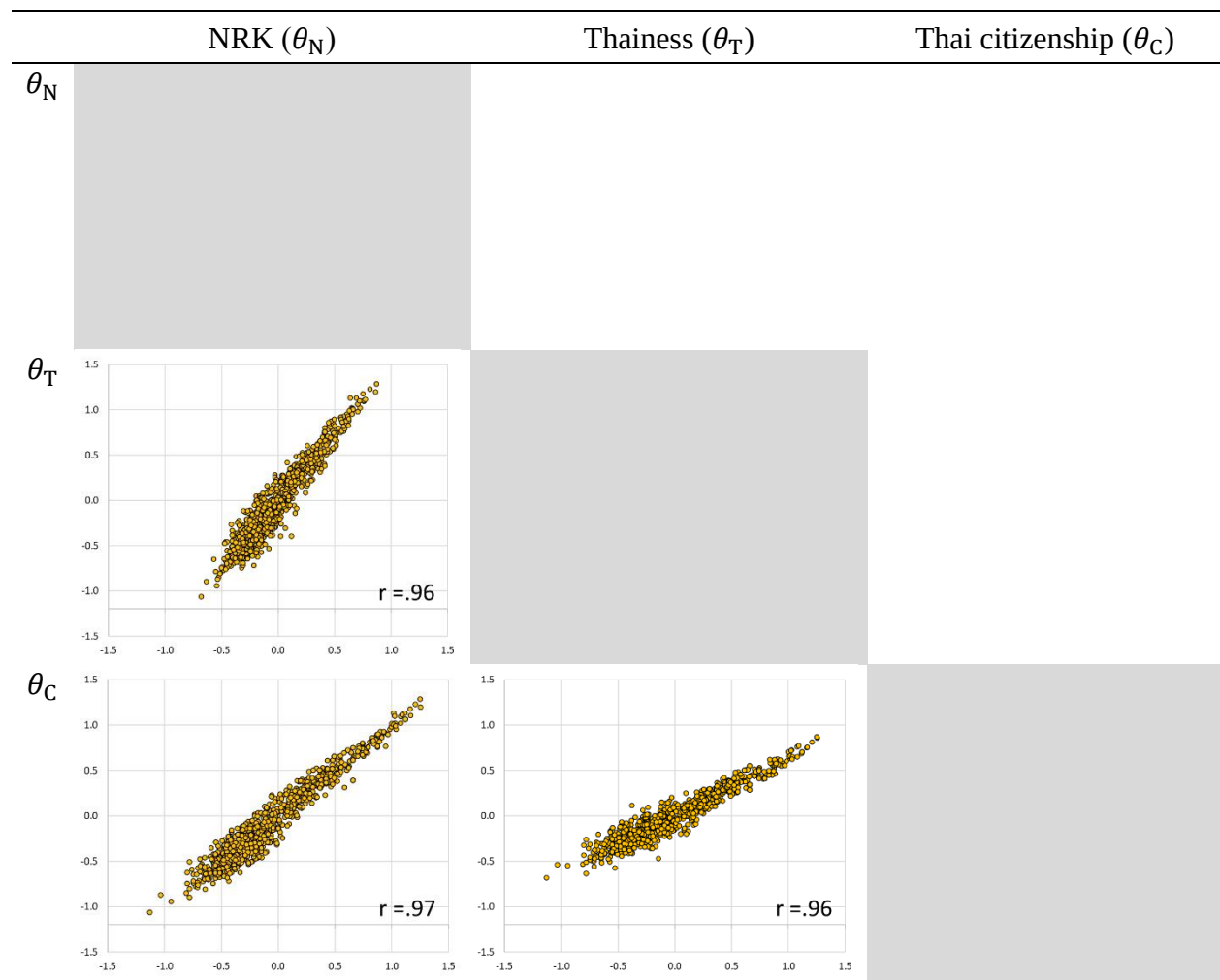


Figure 3.12 *Correlation matrix of student's TCA attributes (total sample)*

The correlation matrix in Figure 3.12 shows that the same finding holds for the total sample to provide high correlation coefficients between .96 to .97 when estimating the student attribute parameters using the multidimensional model. The positive correlations support the underlying theory of TCA construct map. When students possess a high level of TCA, their response scores are high, then the student attribute estimates obtained from the multidimensional model are also high.

The Wright maps in Figures 3.13 to 3.15 show that after the revision, there are remaining 52 TCA items including 16 items for the NRK, 18 items for the Thainess, and 18 items for the Thai Citizenship dimensions, and there are four TCA levels with three item steps.

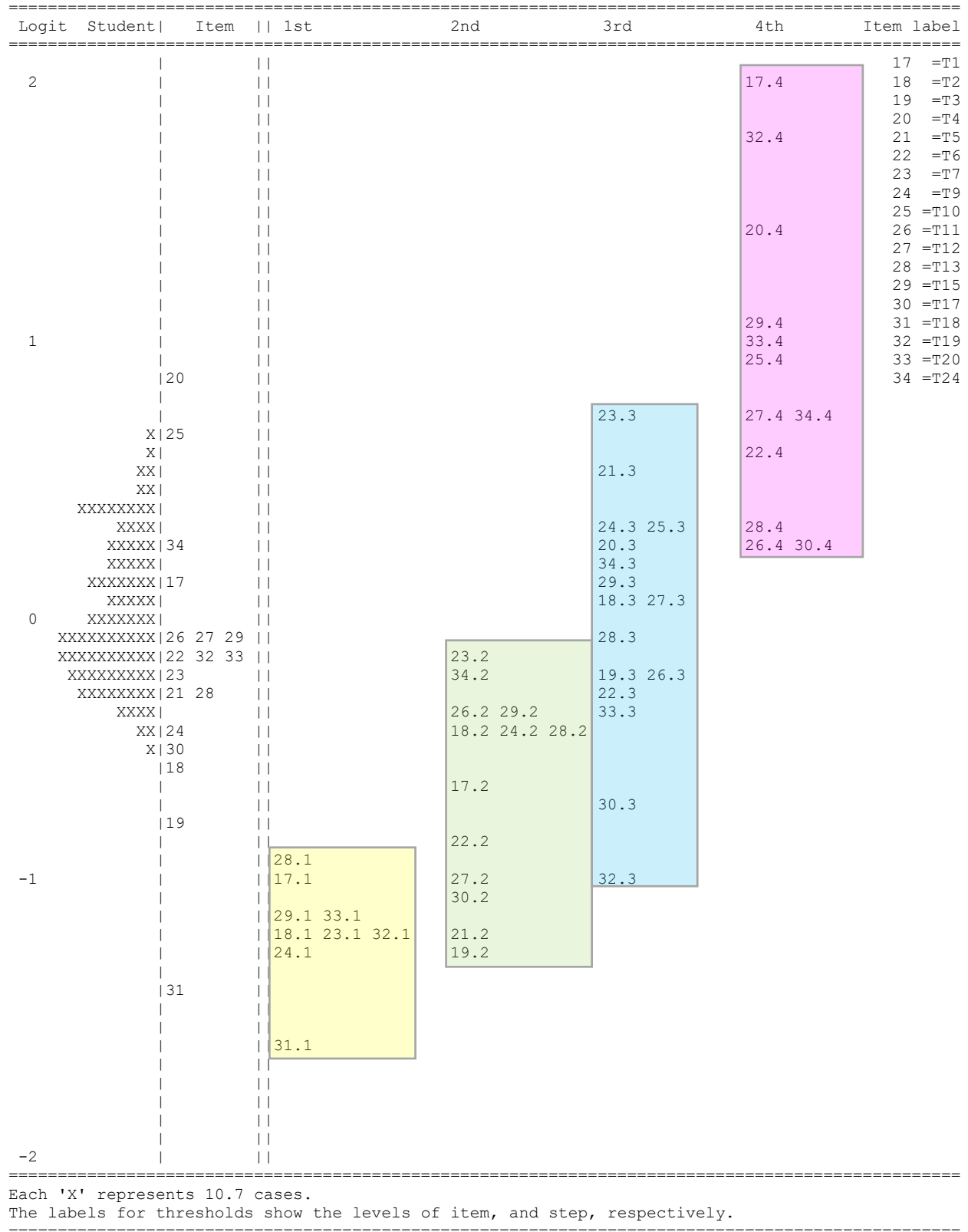


Figure 3.14 The Wright map of Thainess dimension (total sample)

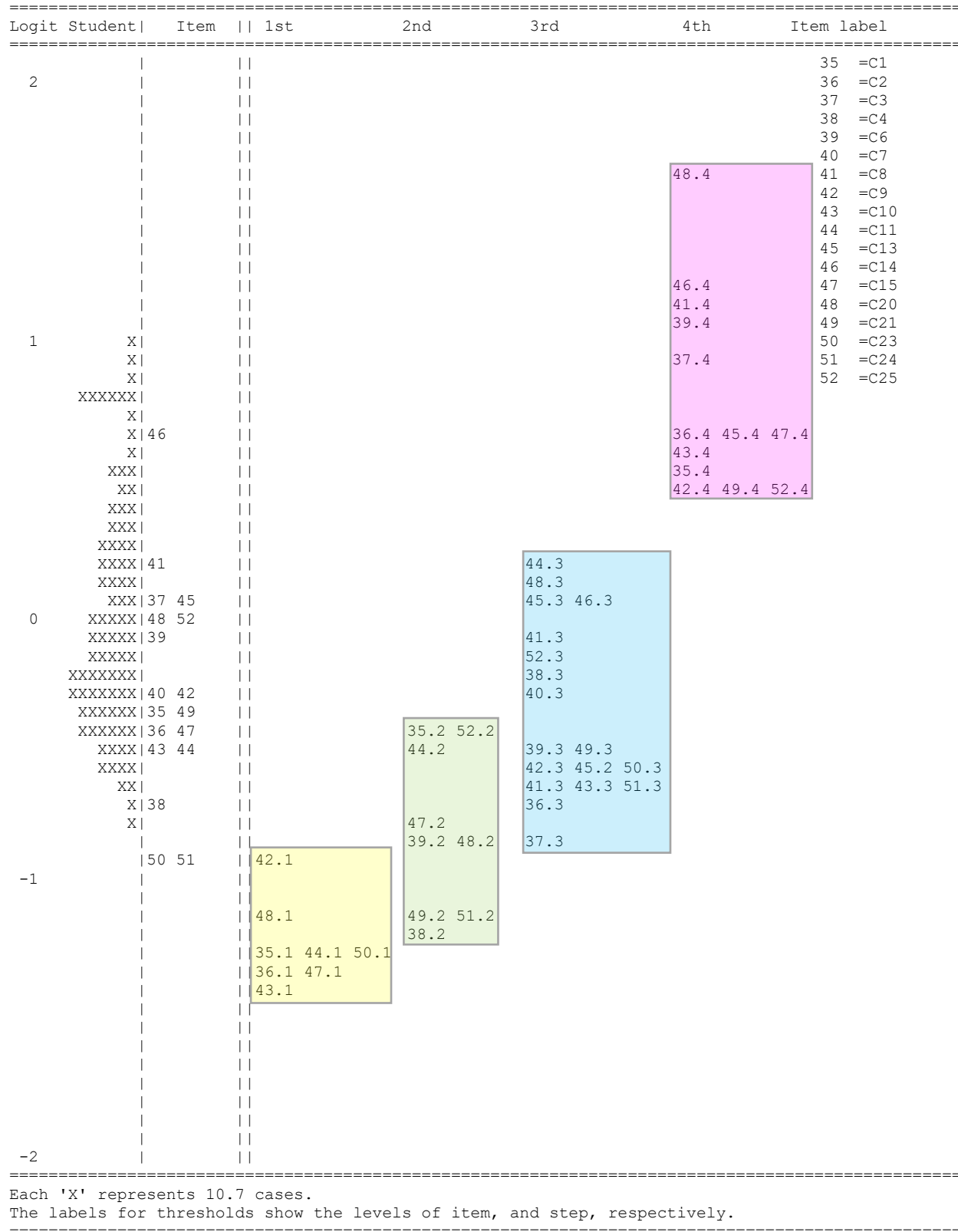


Figure 3.15 The Wright map of Thai Citizenship dimension (total sample)

Table 3.6 *The mean locations trajectories of student attributes in each level of TCA (total sample)*

Item	Level of NRK (score)					Item	Level of Thainess (score)					Item	Level of Thai Citizenship (score)				
	RC	CP	AC	VL	EP		RC	CP	AC	VL	EP		RC	CP	AC	VL	EP
	(0)	(1)	(2)	(3)	(4)		(0)	(1)	(2)	(3)	(4)		(0)	(1)	(2)	(3)	(4)
N1	-.35	-.30		.17	.16	T1	-.22	-.19	.05		.08	C1	-.36	-.14	-.04		.30
N2	-.41	-.27	-.08	.09	.33	T2	-.23	-.09	-.09	.14		C2	-.37	-.31		.17	.07
N3	-.39	-.39	-.26	.16	.31	T3	-.23		.02	.09		C3	-.26			-.05	.40
N4	-.35	-.19	-.19	-.07		T4	-.15			-.03	-.09	C4	-.29		-.25	-.08	
N5		-.36	-.31	-.10	-.07	T5	-.18		-.10	-.08		C6	-.33		-.24	-.11	-.04
N6	-.36	-.29		-.26	.06	T6	-.17		-.19	-.02	-.08	C7	-.18			-.16	
N7	-.43	-.31	-.20		-.07	T7	-.16	-.06	-.17	-.04		C8	-.32		-.24	-.02	.01
N10		-.32	-.27	-.12	-.12	T9	-.23	-.17	-.07	-.05		C9	-.35	-.30		-.06	-.12
N12				-.18	-.16	T10			-.13	-.07	-.05	C10	-.31	-.30		-.08	-.14
N14	-.40		-.08	.38	.38	T11		-.19	-.13	-.09	-.01	C11	-.29	-.26	-.17	-.06	
N15	-.29	-.19		.27	.48	T12	-.08		-.15	-.10	-.04	C13		-.17	-.29	-.09	-.09
N16	-.26	-.07		.49	.05	T13	-.22	-.20	-.17	.01	.02	C14	-.16			.50	.51
N17	-.24	.22		.33		T15	-.21	-.17	-.16	.00	.03	C15	-.30	-.40		.14	.57
N18	-.36		-.11	.15	.49	T17	-.26		-.12	.02	.34	C20	-.41	-.38	.17	.36	.38
N19	-.32	-.12		.20	.43	T18	-.20	-.20				C21		-.34	-.24	.08	.60
N20			-.18	-.06	.47	T19	-.26	-.17		.16	.29	C23	-.35	-.05		.37	
						T20	-.20	-.14		.21	.36	C24	-.29		-.03	.33	
						T24		.17	.08	-.02	.25	C25		-.20	-.17	.20	.57
Mean	-.35	-.22	-.19	.10	.20	Mean	-.20	-.13	-.09	.00	.09	Mean	-.30	-.26	-.15	.09	.23

Comparing to the Wright maps from the previous revision, the same item steps locate closer to each other. There are overlapping locations between the adjacent item steps across three dimensions because each item step contains a degree of uncertainty, usually called standard error. Thus, it makes sense to find such overlapping locations of item steps (and average item parameters) in any measurement. Moreover, there is no item step locations overlapping across more than steps; for example, a first item step location overlaps with a second item step and a third item step. These results are good evidence to support the appropriateness of the item deletion to improve the accuracy of TCA assessment. The results as validity evidence based on internal structure support the accuracy of ordering TCA levels and the plausibility of TCA score interpretation according to the representationalism and pragmatism.

The Wright maps in Figures 3.13 to 3.15 show the connections between the latent NRK, Thainess, and Thai Citizenship attributes and the locations of item steps and student attributes on the logits scales. Thus, the Wright maps connect the contents of items (the content of situations and options) to the content of NRK, Thainess, and Thai Citizenship attributes (the description of student in each level of construct map). The range of item steps locations in each dimension covers the range of student attribute locations; that is the content of items and the content of attribute is matched. This result, thus, as validity evidence based on test content supports the relationship between the content of items and the content of attributes in give the definition of TCA according to the operational perspectives

Table 3.6 show the results when using the total sample, most items show that students with a high TCA level tend to have a high score on each item. This trend relates directly to the developmental continuum structure situated in the three construct maps and the logits score of the Wright maps. However, some items highlighted in blue, yellow, and pink for the NRK, Thainess, and Thai Citizenship dimensions, respectively, show the need of another revision in the future when the increase of student's response scores does not correspond to the increase of student's attribute levels. Considering the mean of mean attribute locations across items in each dimension, overall, there are increasing trends in student attributes from the lowest to the highest level. Overall, these results can be used as validity evidence to support the accuracy of ordering TCA levels. Moreover, the results regarding the group of item steps and the increase of mean attribute locations provide information to develop a guideline for improving the deleted items in the future.

3.10 Score Interpretation

The item step parameters of the NRK, Thainess, and Thai Citizenship dimensions were transformed using the delta dimensional alignment technique (Feuerstahler & Wilson, 2019; Schwartz, 2012) in order to make a reliable comparison of their scores across the dimensions. Table 3.7 show descriptive statistics of EAP student attribute estimates from the unaligned and aligned multidimensional models. The means of EAP student attributes across the dimensions from the aligned model is not constraint to zero, in contrast to the ones from the unaligned model. That means the direct comparison between dimensions can be made in the TCA measurement.

Table 3.7 *Descriptive statistics of student attributes from the dimension alignment analysis*

Dimension	The unaligned model (n=980)				The aligned model (n=980)			
	Mean	SD	Min	Max	Mean	SD	Min	Max
NRK	.00	.46	-1.06	1.29	.04	.46	-1.02	1.30
Thainess	.00	.29	-.68	.87	.03	.31	-.68	1.00
Thai citizenship	.00	.45	-1.13	1.25	.02	.46	-1.10	1.27

In order to make the interpretation of the TCA scores in each dimension, the transformed item steps obtained from the delta dimensional alignment analysis were used to set cut points and criterion zones for discriminating student's attributes in each TCA levels. In Table 3.8, item steps parameters of each item increase from the lowest level to the highest level in each dimension; thus, the means of item steps also increase from the lowest to highest level.

Table 3.8 *Item thresholds statistics*

NRK					Thainess					Thai Citizenship				
Item	1st	2nd	3rd	4th	Item	1st	2nd	3rd	4th	Item	1st	2nd	3rd	4th
N1	-1.07		-.46	1.22	T1	-.98	-.62		1.98	C1	-1.21	-.43		.50
N2	-.87	-.45	.09	.28	T2	-1.21	-.45	.07		C2	-1.32		-.67	.64
N3	-1.12	-.72	-.32	1.25	T3		-1.27	-.20		C3			-.85	.91
N4	-1.30	-.31	.24		T4			.29	1.48	C4		-1.15	-.20	
N5		-1.30	-.52	.80	T5		-1.17	.55		C6		-.83	-.49	1.06
N6	-.76		-.29	.18	T6		-.84	-.27	.63	C7			-.31	
N7	-1.54	-.59		-.24	T7	-1.16	-.15	.73		C8		-.63	-.07	1.13
N10		-1.47	-.60	.77	T9	-1.23	-.44	.34		C9	-.92		-.55	.45
N12				.99	T10			.35	.98	C10	-1.39		-.65	.62
N14		-1.00	-.43	.77	T11		-.38	-.18	.30	C11	-1.22	-.48	.21	
N15	-.93		-.48	.76	T12		-.99	.09	.76	C13		-.56	.02	.65
N16	-.94		-.21	2.02	T13	-.93	-.44	-.05	.37	C14			.05	1.17
N17	-.86		-.23		T15	-1.12	-.36	.10	1.12	C15	-1.30		-.73	.63
N18		-1.59	-.61	.52	T17		-1.05	-.73	.27	C20	-1.07	-.84	.11	1.58
N19	-1.39		-.51	.27	T18	-1.58				C21		-1.08	-.50	.44
N20			-.62	-.06	T19	-1.20		-1.01	1.83	C23	-1.23		-.55	
					T20	-1.13		-.35	1.03	C24		-1.12	-.60	
					T24		-.23	.20	.77	C25		-.43	-.15	.48
Mean	-1.08	-.93	-.35	.68	Mean	-1.17	-.65	0.00	.96	Mean	-1.21	-.76	-.35	.79

In order to create the TCA criterion zone, the first, second, third, and fourth cut points were computed by averaging the first, second, third, and fourth item thresholds, respectively, for each dimension as shown in Table 3.8. Figure 3.16 shows how the criterion zones of each dimension relate to their latent attribute scales and the TCA levels. There are five criterion zones in each dimension: the RC zones refer to the Receiving level, the CP zones refer to the Complying level, the AC zones refer to the Accepting level, the VL zones refer the Valuing level, and the EP zones refer to the Expanding level. Students with the same attribute estimates across the dimensions can be interpreted differently. For example, the -.75 logits fall in the AC zone for the NRK dimension, but in the CP zones for the other two. This student can be interpreted that they are likely to possess the NRK attribute at the Accepting level, and likely to possess the Thainess and Thai citizenship at the Complying level. This result show validity evidence based on testing consequences to support the logic of interpreting TCA scores associated with the TCA levels according to the pragmatic perspective.

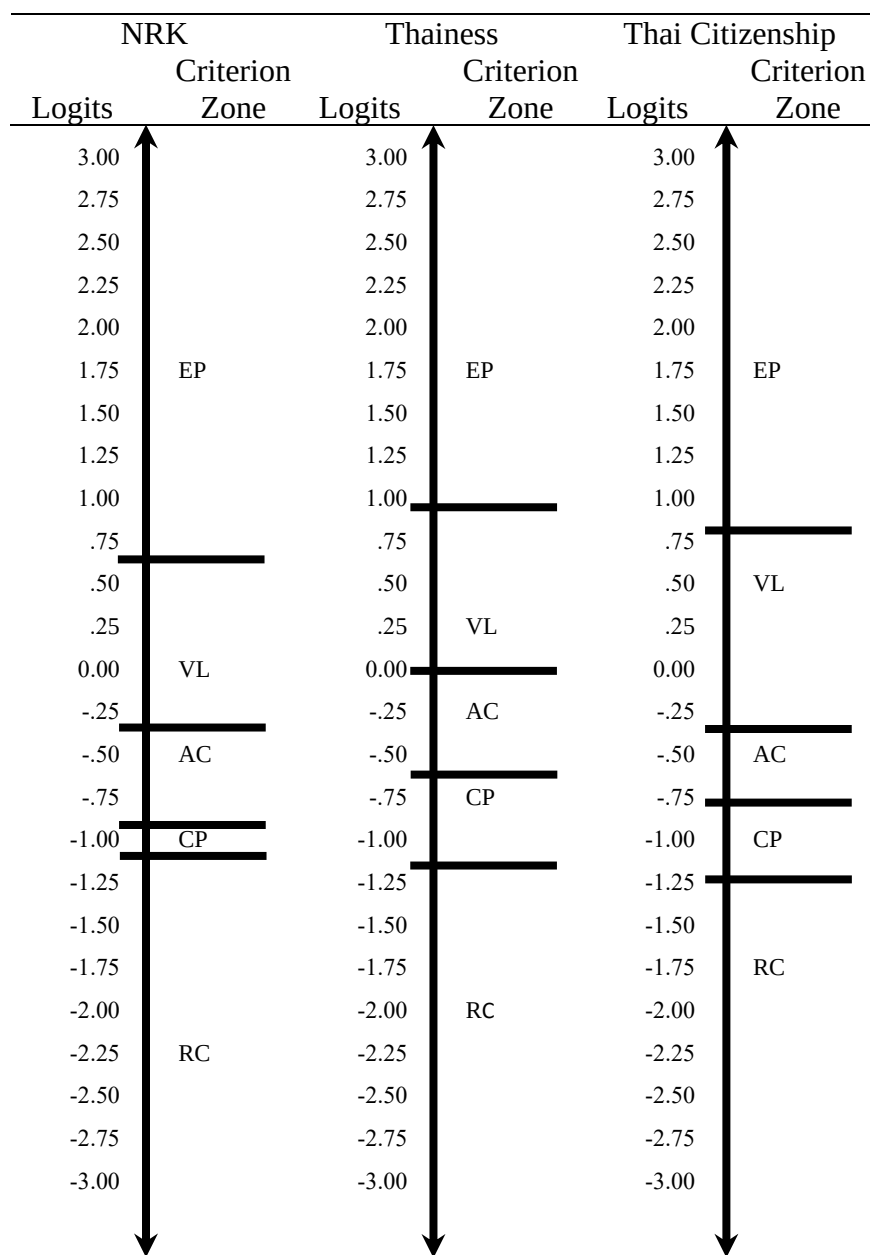


Figure 3.16 *The TCA criterion zone*

Chapter 4

Conclusion

The main purpose of this study is to demonstrate the application of the construct modeling approach (Wilson, 2005) as a measurement design framework with respect to three measurement philosophies—operationalism, representationalism, and pragmatism—and to different types of validity evidence. The assessment of Thai Community Attachment (TCA) for grade 9th students is used to show how the framework works.

The construct modeling approach plays an important role to connecting the measurement philosophies to validity of measurement results. In the TCA assessment, the TCA construct map, which is based on the representational and operational perspectives is the central connector among the TCA attributes. The five-level developmental structure, ranging from Receiving, as the lowest level, to Complying, Accepting, Valuing, and then Expanding, as the highest level, and its relevant description of student attributes who possess different amount of the attributes, reflects the definition of the attributes according to the representational perspective. The three dimensional structure, including National, Religion, and King (NRK), Thainess, and Thai Citizenship content domains, reflects the definition of the attributes according to the operational perspective. That is, the two philosophies are blend together in form of the NRK, Thainess, and Thai Citizenship construct maps.

The three construct maps, which are the most important components of the framework are used in every step of measurement procedure; which means that the representationalist and operationalist perspectives are strongly connected with the measurement procedures. The procedures regarding the TCA item development, student's response scoring, parameter estimation, and score validation, as well as the gathering of validity evidence are systematically connected to the construct maps. At the end of measurement procedure, the TCA scores are interpreted in relation to the levels of construct map; that means the representationalism and pragmatism connect with the TCA score interpretation. When the revision of the TCA assessment is needed, the construct modeling approach provides both guidelines and framing for evidence to support.

Overall, the TCA measurement results show that the construct modeling approach is a useful framework for designing measurements that systematically integrates the measurement foundations, measurement procedures, and measurement results.

References

- Adams, R. J., & Khoo, S-T. (1996). *Quest: The interactive test analysis system*. Australian Council for Educational Research.
- Adams, R. J., Wilson, M., & Wang, W-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1-23. <https://doi.org/10.1177/0146621697211001>
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, 19(6), 716-723. <https://doi.org/10.1109/TAC.1974.1100705>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. <https://www.apa.org/science/programs/testing/standards>
- Anderson, L. W., & Bourke, S. F. (2000). *Assessing affective characteristics in the school* (2nd ed.). Lawrence Erlbaum Associates/Routledge. <https://doi.org/10.4324/9781410605443>
- Bar-Tal, D., & Staub, E. (1997). Introduction: Patriotism: Its scope and meaning. In D. Bar-Tal, & E. Staub (Eds.), *Patriotism in the lives of individuals and nations* (pp. 1-19). Nelson-Hall.
- Ben-Amos, A. (1997). The uses of the past: Patriotism between history and memory. In D. Bar-Tal, & E. Staub (Eds.), *Patriotism in the lives of individuals and nations* (pp. 129-145). Nelson-Hall.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Erlbaum.
- Bradley, W. J., & Schaefer, K. C. (1998). *The uses and misuses of data and models: The mathematization of the human sciences*. SAGE. <https://dx.doi.org/10.4135/9781483348872>
- Bridgeman, P. W. (1927). *The logic of modern physics*. Macmillan.
- Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, 4(1), 87-100.
- Brown, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44(1), 108-132. <https://doi.org/10.1006/jmps.1999.1279>
- Brown, N. J. S., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review*, 23(2), 221-234. <https://doi.org/10.1007/s10648-011-9161-z>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- de Gruijter, D. N. M., & van der Kamp, L. J. Th. (2008). *Statistical test theory for the behavioral sciences*. Chapman & Hall/CRC. <https://doi.org/10.1201/9781584889595>
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41, 417-440. <https://doi.org/10.1146/annurev.ps.41.020190.002221>
- Edwards, M. C., Wirth, R. J., Houts, C. R., & Bodine, A. J. (2015). Three (or four) factor, four (or three) models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 159-182). Routledge. <https://doi.org/10.4324/9781315736013>

- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates; Psychology Press. <https://doi.org/10.4324/9781410605269>
- Feshbach, S., & Sakano, N. (1997). The structure and correlates of attitudes toward one's nation in samples of United State and Japanese college students: A comparative study. In D. Bar-Tal., & E. Staub (Eds.), *Patriotism in the lives of individuals and nations* (pp. 91-107). Nelson-Hall.
- Feuerstahler, L., & Wilson, M. (2019). Scale alignment in between-item multidimensional Rasch models. *Journal of Educational Measurement*, 56(2), 280-301. <https://doi.org/10.1111/jedm.12209>
- Findley, W. G., & Woodruff, A. D. (1964). Book review [Review of the book *Taxonomy of educational objectives, The classification of educational goals, Handbook II: Affective domain*, by D. R. Krathwohl, B. S. Bloom, & B. B. Masia]. *Journal of Educational Measurement*, 1(2), 169-176. <https://www.jstor.org/stable/1433690>
- Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, 4, 26-42. <https://doi.org/10.1037/1040-3590.4.1.26>
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* 5th ed.). Prentice Hall.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27. <https://doi.org/10.1111/j.1745-3992.2004.tb00149.x>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publications.
- Janssen, R., & de Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research*, 34(2), 245-268. <https://doi.org/10.1207/S15327906Mb340205>
- Johnson, G. R. (1997). The evolutionary roots of patriotism. In D. Bar-Tal., & E. Staub (Eds.), *Patriotism in the lives of individuals and nations* (pp. 45-90). Nelson-Hall.
- John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd Ed., pp. 102-138). Guilford Press.
- Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*. National Research Council. <https://bit.ly/3bkSYiN>
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). American Council on Education and Praeger Publisher.
- Kane, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedm.12000>
- Kane, M. T. (2013b). The argument-based approach to validation. *School Psychology Review*, 42(4), 448-457. <https://doi.org/10.1080/02796015.2013.12087465>
- Kane, M. T. (2016). Validity as the evaluation of the claim based on test scores. *Assessment in Education: Principles, Policy & Practice*, 23(2), 309-311. <https://doi.org/10.1080/0969594X.2016.1156645>
- Keeves, J. P., & Masters, G. N. (1999). Introduction. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 1-19). Pergamon.
- Kelman, H. C. (1997). Nationalism, patriotism, and national identity: Social-psychological dimensions. In D. Bar-Tal., & E. Staub (Eds.), *Patriotism in the lives of individuals and nations* (pp. 165-189). Nelson-Hall.

- Kevin, R. R. (1991, March). *Double cross-validation in multiple regression: A method of estimating the stability of results*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Kosterman, R., & Freshbach, S. (1989). Toward a measure of patriotic and nationalistic attitudes. *Political Psychology*, 10(2), 257-274. <https://doi.org/10.2307/3791647>
- Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Volume I: Additive and polynomial representations*. Academic Press.
- Krathwohl, D. R. (1968). The taxonomy of educational objectives-Use of the cognitive and affective domains. In N. E. Gronlund (Ed.), *Readings in measurement and evaluation* (pp. 18-36). Macmillan.
- Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives, The classification of educational goals, Handbook II: Affective domain*. David McKay Company.
- Kyllonen, P. C., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive constructs and their assessment in graduate education: A review. *Educational Assessment*, 10(3), 153-184. https://doi.org/10.1207/s15326977ea1003_2
- Larsen, R. J., & Buss, D. M. (2010). *Personality psychology: Domains of knowledge about human nature* (4th ed.). McGraw-Hill.
- Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions*, 7(4), 328. <https://www.rasch.org/rmt/rmt74m.htm>
- Lipnevich, A. A., MacCann, C., & Roberts, R. D. (2013). Assessing non-cognitive constructs in education: A review of traditional and innovative approaches. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), *The Oxford handbook of child psychological assessment* (pp. 750-772). Oxford University Press.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174. <https://doi.org/10.1007/BF02296272>
- Martin, B. L., & Briggs, L. J. (1986). *The affective and cognitive domains: Integration for instruction and research*. Educational Technology Publications.
- Maxwell, J. A. (2013). *Qualitative research design: An interactive approach* (3rd ed.). SAGE Publications. <https://us.sagepub.com/en-us/nam/qualitative-research-design/book234502>
- Maul, A., Wilson, M., & Irribarra, D. T. (2013). On the conceptual foundations of psychological measurement. *Journal of Physics: Conference Series*, 459. <https://doi.org/10.1088/1742-6596/459/1/012008>
- Maul, A., Irribarra, D. T., & Wilson, M. (2016). On the conceptual foundations of psychological measurement. *Measurement*, 79, 311-320. <https://doi.org/10.1016/j.measurement.2015.11.001>
- McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain*. Springer. <https://doi.org/10.1007/978-1-4614-7135-6>
- McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2), 175-215. <https://doi.org/10.1111/j.1467-6494.1992.tb00970.x>
- Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis, and interpretation of psychological and educational tests*. Eleven International Publishing.
- Messick, S. (1979). Potential uses of noncognitive measurement in education. *Journal of Educational Psychology*, 71(3), 281-292. <https://doi.org/10.1037/0022-0663.71.3.281>

- Ministry of Education of Thailand. (2008). *The basic education core curriculum B.E. 2551 (A.D. 2008)*. Kurusapa Lardpao Printing. <http://www.ipst.ac.th/images/2017/CoreCurriculum2551/EN.pdf>
- Ministry of Education of Thailand. (2014a). *Āib āxng khā niym h̄lāk kh̄xng khn thiṃ* [The twelve core values for Thai people]. <https://www.moe.go.th/websm/2014/sep/198.html>
- Ministry of Education of Thailand. (2014b). *Næwthāng kār cādkār reīyn rû rāywichā pheimteim h̄nāthī phlmeūxng* [The guideline of additional core course of Thai citizenship education]. The Office of the Basic Education Commission. <https://bit.ly/3lD0o5B>
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2), 246-268. <https://doi.org/10.1037/0033-295X.102.2.246>
- Mischel, W., & Shoda, Y. (2008). Toward a unified theory of personality: Integrating dispositions and processing dynamics within the cognitive-affective processing system. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 208-241). Guilford Press.
- Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, 33(4), 379-416. <https://doi.org/10.1111/j.1745-3984.1996.tb00498.x>
- Moseley D., Baumfield, V., Elliott, J., Gregson, M., Higgins, S., Miller, J., & Newton, D. (2005). *Framework of thinking: A handbook for teaching and learning*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511489914>
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). National Academies Press. <https://doi.org/10.17226/10019>
- Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. SAGE Publications. <https://us.sagepub.com/en-us/nam/scaling-procedures/book10174>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- OECD. (2015). *Skill for social progress: The power of social and emotional skills*. OECD Publishing. <https://doi.org/10.1787/23078731>
- OECD. (2017). *PISA 2015 technical report*. OECD Publishing. <https://www.oecd.org/pisa/data/2015-technical-report/>
- Office of the National Economic and Social Development Board. (2007). *Sufficiency economy implications and applications*. http://www.nesdb.go.th/Md/book/booksuffwork_eng.pdf
- Paek, I. (2002). *Investigations of differential item functioning: Comparisons among approaches, and extension to a multidimensional context* (Publication No. 3082356). [Doctoral dissertation, University of California, Berkeley]. ProQuest Dissertations & Theses Global.
- Primoratz, I. (2007). Patriotism and morality: Mapping the terrain. In I. Primoratz & A. Pavkovic (Eds.), *Patriotism: Philosophical and political perspectives* (pp. 17-35). Ashgate.
- Primoratz, I. (2017). Patriotism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2017 ed.). Stanford University. <https://plato.stanford.edu/entries/patriotism/>
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press. (Original work published 1960)

- Reykowski, J. (1997). Patriotism and the collective system of meanings. In D. Bar-Tal., & E. Staub (Eds.), *Patriotism in the lives of individuals and nations* (pp. 108-128). Nelson-Hall.
- Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality*, 43(2), 137-145. <https://doi.org/10.1016/J.jrp.2008.12.015>
- Rosenberg, M. J., & Hovland, C. I. (1960). Cognitive, affective, and behavioral components of attitudes. In C. I. Hovland & M. J. Rosenberg (Eds.), *Attitude organization and change: An analysis of consistency among attitude components* (pp. 1-14). Yale University Press.
- Schermelleh-Engel, K., Moosbrugger, H., & Muller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74. <https://www.scienceopen.com/document?vid=4c30cf93-535c-492f-bdf5-19a40a6af21b>
- Schwartz, R. A. (2012). *The development and psychometric modeling of an embedded assessment for a data modeling and statistical reasoning learning progression* (Publication No. 3555912). [Doctoral dissertation, University of California, Berkeley]. ProQuest Dissertations & Theses Global.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464. <https://www.jstor.org/stable/2958889>
- Secretariat of the House of Representatives of Thailand. (2017). *Constitution of the Kingdom of Thailand 2017: Synopsis*. <https://bit.ly/3jUxY69>
- Stevens, S. S. (1946). On The theory of scales of measurement. *Science*, 103(2684), 677-680. <https://www.jstor.org/stable/1671815>
- Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). John Wiley & Sons.
- Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, 161(3844), 849-856. <http://www.jstor.org/stable/1724851>
- Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. John Wiley & Sons.
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6). <https://doi.org/10.1007/s11165-016-9602-2>
- Tal, E. (2017). Measurement in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Fall 2017 ed.)*. Stanford University. <https://plato.stanford.edu/archives/fall2017/entries/measurement-science/>
- Thomson, W. (1889). Electrical units of measurement. In W. Thomson (Ed.), *Popular Lectures and Addresses: Vol. 1, Constitution of matter* (pp. 73-136). Cambridge University Press. <https://doi.org/10.1017/CBO9780511997242.006>
- Trendler, G. (2019). Measurability, systematic error, and the replication crisis: A reply to Michell (2019) and Krantz and Wallsten (2019). *Theory & Psychology*, 29(1), 144-151. <https://doi.org/10.1177/0959354318824414>
- Wang, W.-C., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, Jr & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 139-156). Ablex.

- Werner, C., & Schermelleh-Engel, K. (2010). *Introduction to Structural Equation Modeling with LISREL*. Retrieved from http://www.psychologie.uzh.ch/dam/jcr:ffffffffff-b371-2797-0000-00000fda8f29/chisquare_diff_en.pdf.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates; Routledge. <https://doi.org/10.4324/9781410611697>
- Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6), 716-730. <https://doi.org/10.1002/tea.20318>
- Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, 60(2), 181-198. <https://doi.org/10.1007/BF02301412>
- Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling. *Health Education Research*, 21(Supplement 1), i4-i18. <https://doi.org/10.1093/her/cyl108>
- Wilson, M., Bejar, I., Scalise, K., Templin, J., William, D., & Irribarra, D. T. (2012). Perspectives on methodological issues. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills*, (pp. 67-142). https://doi.org/10.1007/978-94-007-2324-5_3
- Wilson, M., & Masters, G. N. (1993). The partial credit model and null categories. *Psychometrika*, 58(1), 87-99. <https://doi.org/10.1007/BF02294473>
- Wilson, M., & Moore, S. (2012). An explanative modeling approach to measurement of reading comprehension. In J. P. Sabatini, T. O'Reilly, & E. R. Albro (Eds.), *Reaching an understanding: Innovations in how we view reading assessment* (pp. 147-168). Rowman & Littlefield Education,
- Wilson, M., & Wang, W-C. (1995). Complex composites: Issues that arise in combining different models of assessment. *Applied Psychological Measurement*, 19 (1), 51-71. <https://doi.org/10.1177/014662169501900107>
- Wright, B. D. (1984). Additivity in psychological measurement. In E. E. Roskam (Ed.), *Measurement and personality assessment* (pp. 101-111). Elsevier Science.
- Wright, B. D., & Masters, G. N. (1990). Computation of outfit and infit statistics. *Rasch Measurement Transactions*, 3(4), 84-85. <https://www.rasch.org/rmt/rmt34e.htm>
- Wright, B. D., & Tennant, A. (1996). Sample size again. *Rasch Measurement Transactions*, 9(4), 468. <https://www.rasch.org/rmt/rmt94h.htm>
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest version 2.0: generalised item response modeling software [computer software]. ACER Press.