1

INTRODUCTION

Georg Rasch summarized his basic research on measurement in a book entitled *Probabilistic Models for Some Intelligence and Attainment Tests* (1960/1980). Rasch's research led to the development of a new paradigm for measurement in the social sciences. As pointed out by van der Linden (2016), the first chapter of Rasch's book is mandatory reading for anyone seeking to understand the transition from classical test theory (CTT) to item response theory (IRT). Our book focuses on describing Rasch's theoretical and applied contributions to measurement theory and the use of the Rasch model as a framework for solving a set of important measurement problems. We also stress the continuing relevance of Rasch's contributions to modern measurement for the social, behavioral, and health sciences.

What are the key aspects of Rasch's contributions that provided the basis for a paradigm shift in measurement theory? First of all, Rasch recognized that measurement should focus on individuals. In his words, "present day statistical methods are entirely group-centered, so that there is a real need for developing individual-centered statistics" (Rasch, 1961, p. 321). His solution to this problem led him to propose a set of requirements for specific objectivity in the individual-centered measurement:

- The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and
- it should also be independent of which stimuli within the considered class were or might also have been compared.
- Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and
- it should also be independent of which other individuals were also compared on the same or on some other occasion.

(Rasch, 1961, pp. 331–332)

The first two bullet points suggest that item calibrations (stimuli) should be invariant over the persons that are used to obtain the

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

comparisons: person-invariant calibration of items. The last two bullet points suggest that person measurement should be invariant of the particular items (stimuli) that are used to obtain the comparisons: item-invariant measurement of persons.

Second, Rasch proposed a measurement model for achieving item-invariant measurement of persons and person-invariant calibration of items. This contribution is fundamental because it provides the basis for the separation of items and persons. This separation allows for person scores to be independent of the particular items and for item locations (difficulties) to be estimated independently of the particular persons. In Rasch's research related to IRT, he highlights that the probability of a correct response to an item should be a simple function of item difficulty and person ability.

Another important contribution of Rasch measurement theory is that it provides a philosophical approach stressing that measurement requires a strong set of requirements, and that data may or may not fit the Rasch measurement model. This distinction is essential for achieving invariant measurement with real data. The opportunity to refute the model by evaluating model-data fit is a strong characteristic of Rasch's requirements for specific objectivity and the accomplishment of invariant measurement.

This chapter begins with a brief overview of invariant measurement within three research traditions that can be used for classifying measurement theories. We define the role of invariance in each of these traditions and highlight the importance for invariance in measurement. Second, Rasch measurement theory is introduced as a framework for obtaining invariance in the social sciences. The following section provides an overview of the key components that are essential for scale development from the perspective of Rasch measurement theory. Next, we define several major measurement problems that are discussed in this book including the definition of latent variables, evaluation of differential item functioning, examination of the interchangeability of items for person measurement, and creation of performance standards (cut scores) through standard setting. Finally, an overview is provided of the key topics addressed in this book.

1.1 Invariant Measurement

The scientist seeks measures that will stay put while his back is turned.

(Stevens, 1951, p. 21)

Do not copy, post or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

This section provides a brief overview of three research traditions that can be used to organize the major theoretical perspectives that dominate measurement theory: test score, scaling, and structural traditions. We discuss how these traditions relate to invariant measurement. Finally, we introduce how Rasch measurement theory (scaling tradition) is used as a framework for understanding invariant measurement.

Research Traditions in Measurement

In order to understand the evolution of measurement theory and the concept of invariance during the 20th century, it is useful to consider three broad paradigms or research traditions that have dominated the field. The first tradition is a test score tradition that is reflected in CTT. The second tradition is a scaling tradition that is reflected in Rasch measurement theory including IRT in general. The third tradition is a structural tradition that can be represented by structural equation modeling (SEM). There are several key distinctions between measurement theories embedded within each research tradition, and these distinctions influence the perspective on invariance within each tradition. These distinctions are (1) focus of the measurement theory, (2) form of the underlying model, (3) perspective on invariance, (4) overarching goal of measurement theories within each tradition, and (5) visual representations of measurement models. Figure 1.1 summarizes some of the major differences between measurement theories that are embedded within these broader research traditions.

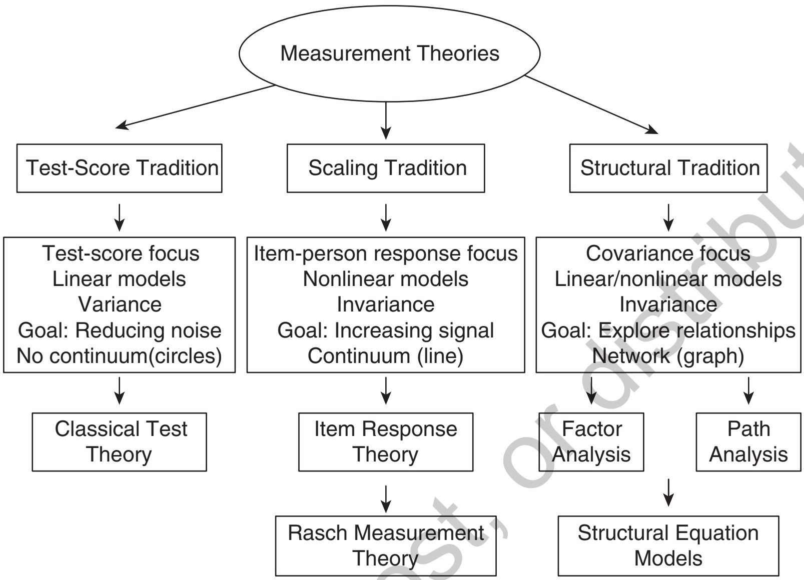
The focus of measurement theories within the test score tradition is on sum scores. The underlying statistical model is linear as shown in the standard decomposition of an observed score (O) into a true score (T) and error score (E): $O = T + E$. The underlying statistical model for observed test scores is linear. The definition of CTT is tautological—the model is assumed to be true by definition (Traub, 1997). Measurement theories in the test score tradition focus on the variance of test scores. The overarching goal is to reduce measurement error and minimize potential sources of construct-irrelevant variance. Overall, the test score tradition focuses on reducing noise which is variant. Visual depictions of measurement theories in the test score tradition include the use of Venn diagrams to illustrate different sources of variance in test scores (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Measurement theories within the scaling tradition focus on detailed responses of each person to a carefully selected set of items. Rasch

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Figure 1.1 Three Traditions in Measurement Theory



measurement theory as described in this book is an example of models in the scaling tradition. The underlying statistical model is nonlinear and probabilistic. Many of the measurement models in the test score tradition are descriptive in nature, while models within the scaling tradition emphasize strict requirements for model-data fit in order to support the inferences-related invariant measurement. The form of the Rasch model is shown later in this chapter, and a detailed description of model-data fit is provided in Chapter 3. Measurement theories in the scaling tradition emphasize the quest for invariant measurement and the creation of scales to "stay put while our backs are turned" (Stevens, 1951, p. 21). Measurement theories in the scaling tradition have the goal of increasing the signal in our measures, as compared to the reduction of noise emphasized in the test score tradition. The visual depictions of measurement theories in the scaling tradition involve the creation of continuums to represent latent variables (constructs) on a line.

The final family of measurement theories we briefly consider here is within the structural tradition. The structural tradition focuses on

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

covariances (or correlations). Both linear and nonlinear models have been used to examine the relationships between observable and latent variables. Measurement invariance related to models within the structural tradition has been extensively discussed by Millsap (2011). The overarching goal of structural models is to explore relationships between latent variables. Both factor analysis and path analysis including the combination of these models within SEM are multivariate (Bollen, 1989). These networks of variables are typically represented as path diagrams and graphs.

A useful mnemonic device to remember distinctions between the traditions is that the test score tradition relates to circles, the scaling tradition relates to lines, and the structural tradition relates to graphs. The first two traditions were described earlier by Engelhard (2013), while the third tradition is a new addition to the conceptual framework for grouping measurement theories into research traditions. This book focuses on Rasch measurement theory that is within the scaling tradition and the views of invariant measurement from this research tradition.

Invariant Measurement and the Scaling Tradition

Turning now to invariant measurement within the scaling tradition, the basic requirements have been widely recognized by measurement theorists throughout the 20th century. In one of the first books on measurement, Thorndike (1904) outlined the basic ideas of invariant measurement (Engelhard, 2013). It is important to recognize that invariant measurement has two aspects: item-invariant person measurement and person-invariant item calibration. Figure 1.2 lists the requirements for invariant measurement related to Rasch measurement theory. Requirements 1 and 2 imply item-invariant person measurement. Requirements 3 and 4 imply person-invariant item calibration, while the last requirement implies that invariant measures should be unidimensional. Each of these requirements is described in more detail later in our book.

It is of critical importance to recognize that the requirements of invariant measurement are a set of hypotheses that must be verified with a variety of model-data fit procedures. In this book, we stress the use of residual analyses and fit indices for evaluating invariance for person measurement, item calibration, and unidimensionality of the scale. Invariant measurement is not automatically obtained by the

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Figure 1.2 Five Requirements for Invariant Measurement

Five Requirements for Invariant Measurement

Item-invariant measurement of persons

1. The measurement of persons must be independent of the particular items that happen to be used for the measuring.
2. A more able person must always have a better chance of success on an item than a less able person (Non-crossing person response functions).

Person-invariant calibration of test items

3. The calibration of the items must be independent of the particular persons used for calibration.
4. Any person must have a better chance of success on an easy item than on a more difficult item (Non-crossing item response functions).

Unidimensional scale

5. Items and persons must be simultaneously located on a single underlying latent variable (Wright map).

use of Rasch models—the requirements must be evaluated with specific samples from the target population. Chapter 3 on evaluating a Rasch scale describes these issues in greater detail.

1.2 Rasch Measurement Theory

> *The psychometric methods introduced in [Rasch's book] … embody the essential principles of measurement itself, the principles on which objectivity and reproducibility, indeed all scientific knowledge are based.*
>
> (Wright, 1980, p. xix)

Rasch measurement theory provides a basis for meeting the requirements of invariant measurement as described in Figure 1.2. The focus of our book is on Rasch measurement theory, but it should be recognized that other measurement models within the scaling tradition, such as IRT, can also meet some of these requirements for invariant measurement. Rasch measurement theory is based on a simple idea about what happens when a single person encounters a single test item (Rasch, 1960/1980). Rasch viewed the probability of a person getting an

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

item correct (or endorsing an item) as a function of the person's ability measure and the difficulty of the item.

Rasch measurement models are fundamentally based on the concept of comparisons. In fact, Rasch (1977) argued that all scientific statements "deal with comparisons, and the comparisons should be objective" (p. 68). Abelson (1995) stressed in his book on *Statistics as Principled Argument* that comparisons are crucial for all meaningful and useful statistical analyses. Within the context of measurement theory, Andrich (1988) highlighted that the "demand for invariance of comparisons across a scale that is unidimensional is paramount" (p. 43). The task for scientists and measurement theorists is to define what is meant by invariant comparisons and objectivity. Rasch proposed a model based on the principle of specific objectivity that provides a framework for these invariant comparisons, namely the Rasch model. The Rasch model for dichotomous responses can be written as:

$$\phi_{ni1} = \frac{\exp(\theta_n - \delta_{i1})}{1 + \exp(\theta_n - \delta_{i1})} \tag{1.1}$$

The Rasch model in Equation 1.1 can be viewed as an operating characteristic function that relates the differences between locations of persons (θ) and items (δ) on a latent variable to the probability of a positive response on a dichotomous item (e.g., correct answer on a multiple-choice item, "yes" response to a survey item). This distance reflects a comparison between each person's ability and item difficulty that predicts a probability of an affirmative answer by a person on each item. This book illustrates the problem-solving capabilities of this elegant approach to measurement proposed by Rasch (1960/1980).

Many current psychometric texts provide general descriptions of IRT models, including the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models (Baker & Kim, 2004). Equation 1.2 shows a general expression for a 3PL model that includes three item parameters—difficulty, discrimination, and pseudo-guessing. The pseudo-guessing parameter defines a lower asymptote for the probability response curve and the minimum probability to get an item correct. If the pseudo-guessing effect is not considered (c_i = 0), then this expression reduces to a 2PL model that contains item difficulty and discrimination parameters. Item discrimination parameter reflects the slope or the steepness of an item response function. If a_i is a constant and assuming that item response functions are parallel with common slope, then Equation 1.2 is simplified to

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Do not copy, post, or distribute

represent a 1PL model that estimates only item difficulty parameters. From this perspective, the Rasch model can be obtained when $a_i = 1$ and $c_i = 0$:

$$P_{ij}(X = 1) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}, \quad (1.2)$$

where

P_{ij} = probability of correct response,

b_i = item difficulty parameter,

a_i = item discrimination parameter,

c_i = item lower asymptote (pseudo-guessing) parameter, and

θ_j = latent proficiency of person.

Within the context of general IRT models, some researchers treat the 1PL model as nested within a more general IRT model, and that the 1PL is "equivalent to the well-known Rasch model" (Raykov & Marcoulides, 2011, p. 294). We would like to stress that although the Rasch model can be defined using general IRT model forms, this perspective on the Rasch model does not consider the full philosophical basis of Rasch's contribution as a specific approach to measurement in the social, behavioral, and health sciences. This perspective on the Rasch model as a special case of a more general model creates a specific set of evaluation criteria that stresses how well the models fit data—since including more parameters in a model generally provides better fit, then this perspective, ceteris paribus, privileges more complex IRT models.

In this book, we argue that the fundamental principles and requirements of measurement should come first, and that measurement is the proactive creation of a scale that defines a latent variable based on requirements of invariant measurement—the goal is not to find the best statistical model to fit the data, but the psychometric goal of creating meaningful and useful scales based on a strict set of requirements for high-quality measures. As pointed out by Wright (1980):

> *If the data cannot be brought into an order which fits with a measurement model, then we will not be able to use these data for measurement, no matter what we do to them. The requirement for results that we are willing to call measures are specified in the model*

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

> *and if the data cannot be managed by the models, then the data cannot be used to calibrate items or measure persons.*
>
> (p. 193)

Andrich (1989) provided an excellent discussion of these issues including an insightful discussion of the distinction between requirements and assumptions in measurement theory.

Rasch measurement theory views model-data fit in terms of a specific set of data meeting the requirements of the model. In order to have the desirable characteristics of invariant measurement, it is essential to have good fit to the model. The basic distinction is that Rasch measurement theory starts with the requirements of the model, while IRT more generally evaluates the success of the models in terms of how well a particular data set is reproduced. We provide a brief comparison between Rasch and 2PL models to show the importance of meeting the requirements of invariant measurement.

The person response functions of four persons are displayed in Figure 1.3 in order to illustrate the importance of item-invariant measurement of persons (Requirements 1 and 2 for invariant measurement). Based on the Rasch model, Persons A, B, C, and D have location measures at −1.00, −0.50, 0.50, and 1.00 logits, respectively. Person A is viewed as having the lowest level of proficiency, while Person D has the highest proficiency on the latent variable. With the 2PL model, slope parameters for persons (Engelhard & Perkins, 2011) are included which are set to 0.30 logits for Person A, 1.60 logits for Person B, 0.80 logits for Person C, and 1.50 logits for Person D.

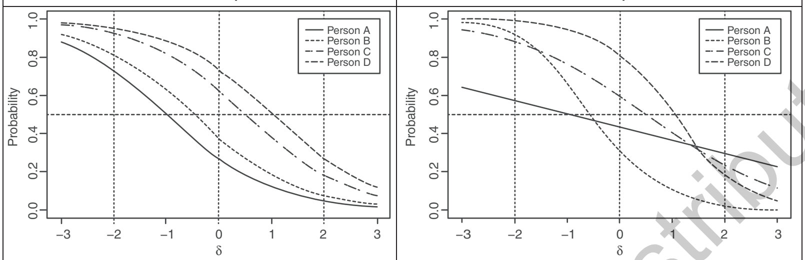
As we can see in Figure 1.3, the person response functions based on the Rasch model are noncrossing (Panel A); however, the 2PL model yields crossing person response curves (Panel E). Next, probabilities of a correct response are computed for answering three items with difficulty values of −2.00, 0.00, and 2.00 logits separately. Based on the Rasch model, the most proficient Person D has the highest probability of answering each item correct, and Person A who is the least proficient person has the lowest probability of correct on every item. The probability of getting an item correct is higher for Person C who has medium high proficiency than for Person B who has medium low proficiency (Panel B). Based on the 2PL model, the orders of the probabilities are not consistent for different items (Panel F). For instance, the least proficient Person A (p = 0.29) had higher probability than Person D (p = 0.18) in answering Item 3 correctly. Finally, we order the persons based on their probabilities of answering each item correct.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Figure 1.3 Examining Item-Invariant Measurement of Persons With Person Response Functions

A. Rasch Model Response Functions | E. 2PL Model Response Functions



B. Rasch Model: Probabilities

Rasch Person Order	Items 1 Easy	2 Medium	3 Hard
D (High)	.96	.73	.27
C (Medium High)	.92	.62	.18
B (Medium Low)	.82	.38	.08
A (Low)	.73	.27	.05

F. 2PL Model: Probabilities

Person Order	Items 1 Easy	2 Medium	3 Hard
(High)	.99	.82	.18
(Medium High)	.88	.60	.23
(Medium Low)	.92	.31	.02
(Low)	.57	.43	.29

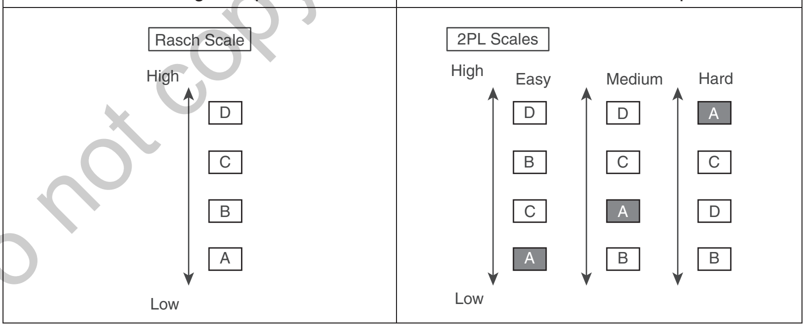
C. Invariant person order

Rasch Person Order	Items 1 Easy	2 Medium	3 Hard
D (High)	D	D	D
C (Medium High)	C	C	C
B (Medium Low)	B	B	B
A (Low)	A	A	A

G. Variant person order

Person Order	Items 1 Easy	2 Medium	3 Hard
(High)	D	D	A
(Medium High)	B	C	C
(Medium Low)	C	A	D
(Low)	A	B	B

D. Wright Map | H. 2PL Person Maps



This explicitly shows the invariant order of persons based on the Rasch model (Panels C and D), but inconsistent patterns obtained with a 2PL model (Panels G and H).

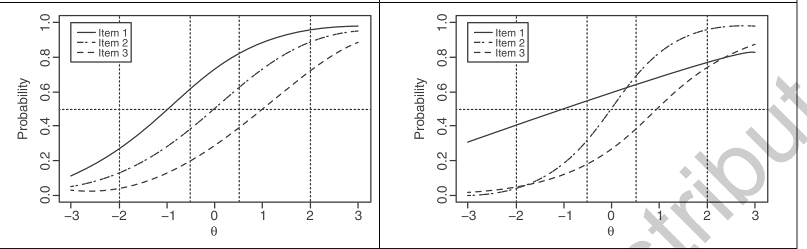
Figure 1.4 illustrates the importance of person-invariant calibration of items (Requirements 3 and 4 for invariant measurement). The probability response functions for three items are shown. The three

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Figure 1.4 Examining Person-Invariant Calibration of Items With Item Response Functions

A. Rasch Model Response Functions | E. 2PL Model Response Functions



B. Rasch Model: Probabilities

Rasch Item Order	Persons: A Low	B Medium Low	C Medium High	D High
3 (Hard)	.05	.18	.38	.73
2 (Medium)	.12	.38	.62	.88
1 (Easy)	.27	.62	.82	.95

F. 2PL Model: Probabilities

Item Order	Persons: A Low	B Medium Low	C Medium High	D High
(Hard)	.05	.18	.38	.73
(Medium)	.04	.31	.69	.96
(Easy)	.40	.55	.65	.77

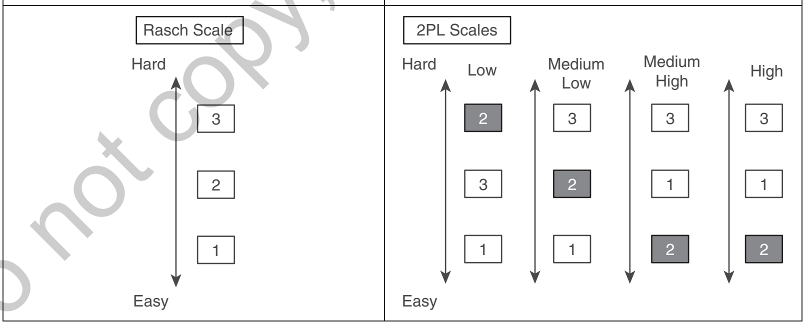
C. Invariant item order

Rasch Item Order	Persons: A Low	B Medium Low	C Medium High	D High
3 (Hard)	3	3	3	3
2 (Medium)	2	2	2	2
1 (Easy)	1	1	1	1

G. Variant item order

Item Order	Persons: A Low	B Medium Low	C Medium High	D High
(Hard)	2	3	3	3
(Medium)	3	2	1	1
(Easy)	1	1	2	2

D. Wright Map | H. 2PL Item Maps



Do not copy, post, or distribute

items based on the Rasch model have difficulty values of −1.00 (Item 1), 0.00 (Item 2), and 1.00 (Item 3) logits. Therefore, Item 1 is the easiest, Item 2 is of medium difficulty, and Item 3 is the hardest. The three items of 2PL model not only have the same difficulty parameters as the Rasch items but also have discrimination parameters that vary as follows: 0.40 (Item 1), 1.60 (Item 2), and 1.00 (Item 3) logits.

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The Rasch model provides noncrossing item response functions (Panel A); however, the 2PL model has different slopes, and this leads to crossing item response functions (Panel E). Second, the probabilities of getting each item correct have been calculated for four persons—A, B, C, and D with different locations on the latent variable scale which are −2.00, −0.50, 0.50, and 2.00 logits, respectively. Based on a Rasch model, every person has the highest probability of answering the easiest item (Item 1) correct and lowest probability of getting the hardest item (Item 3) correct (Panel B). According to the 2PL model, Person D has higher probability to correctly answer Item 2 ($p = 0.96$), which is of medium difficulty than the easiest Item 1 ($p = 0.77$) as shown in Panel F. Based on the probabilities, we see that the order of the items is invariant with the Rasch model (Panels C and D), while the 2PL model yields variant ordering of items that is dependent on where the person is located on the latent variable (Panels G and H).

The final requirement for invariant measurement is unidimensionality (Requirement 5 for invariant measurement). There are a variety of ways to conceptualize dimensionality. We are guided in our work by the views of Louis Guttman. In the 1940s, Guttman (1944, 1950) laid the groundwork for a new technique designed to explore the unidimensionality of a set of test items. According to Guttman (1950):

> *One of the fundamental problems facing research workers … is to determine if the questions asked on a given issue have a single meaning for the respondents. Obviously, if a question means different things to different respondents, then there is no way that the respondents can be ranked … Questions may appear to express a single thought and yet not provide the same kind of stimulus to different people.*
>
> (p. 60)

Guttman scaling can be viewed as an approach for determining whether or not a set of items and a group of persons meet the requirements of unidimensionality (Engelhard, 2008a). As shown in Figures 1.2 and 1.3, the Rasch scale can be viewed as a probabilistic Guttman scale. This is very important as pointed out by Cliff (1983):

> *the Guttman scale is one of the very clearest examples of a good idea in all of psychological measurement. Even with an unsophisticated—but intelligent—consumer of psychometrics, one*

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

has only to show him a perfect scale and the recognition is almost instantaneous, "Yes, that's what I want."

(p. 284)

Guttman scaling provides a deterministic framework for examining dimensionality, while Rasch measurement theory can be viewed as a probabilistic extension of Guttman scaling theory. A scale is unidimensional if the properties of basic ordering of person-invariant item calibrations and item-invariant measurement are met in the data set.

One of the major implications of meeting the requirements of invariant measurement is that researchers can create a unidimensional scale. With the Rasch model, both items and persons are simultaneously located on an underlying latent scale. As shown earlier, the 2PL model implies different location orderings of items and persons with different implications for the interpretation, meaning, and use of the scores obtained on a scale.

Later in this book, we demonstrate methods for checking if the requirements of Rasch measurement theory have been achieved. Our approach to model-data fit depends mainly on the use of various types of residual analyses to evaluate the requirements of invariant measurement for a particular data set.

1.3 Components of Scale Development Based on Rasch Measurement Theory

The process by which concepts are translated into empirical indices has four steps: an initial imagery of the concept, the specification of dimensions, the selection of observable indicators, and the combination of indicators into indices.

(Lazarsfeld, 1958, p. 109)

One of the most important tasks in the social, behavioral, and health sciences is the development of substantive theories of human behavior based on a well-developed, understandable, and agreed-upon set of concepts. The "initial imagery of the concept" (Lazarsfeld, 1958, p. 109) plays a central role in our theories in the human sciences. As pointed out by Lazarsfeld (1966), "problems of concept formation, of meaning, and of measurement necessarily fuse into each other" (p. 144).

It can be argued that the measurement aspects of social science research are the least developed and understood area of research with

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

great potential for confounding our understanding of human behavior and action. If there is not a clear understanding of units and the meaning of our measures, then it is virtually impossible to have any practical plan for implementing a theory of action based on our theoretical framework. Rasch measurement theory provides a basis for developing a set of stable measures to anchor and critically evaluate the broader theoretical models of human behavior.

Stone, Wright, and Stenner (1999) describe how measurement is made by analogy, and that maps can provide visual representations of latent variables. They highlight how useful other maps, such as rulers (length), clocks (time), and thermometers (temperature), are for communicating an underlying latent variable by analogy. In their words, "Successful item calibration and person measurement produce a map of the variable. The resulting map is no less a ruler than the ones constructed to measure length" (Stone et al., 1999, p. 321).

A Wright map represents a Rasch scale developed to measure a latent variable (e.g., food insecurity, math proficiency, and learning motivation), and it should be viewed as both a set of hypotheses regarding the latent variable and ultimately a validated definition of the latent variable defined by the scale. Wright maps have also been called variable maps and item maps. Wilson (2011) proposed calling them *Wright maps* in order to honor Professor Benjamin D. Wright at the University of Chicago who was a major figure in expanding the development and use of Rasch measurement theory around the world (Wilson & Fisher, 2017).

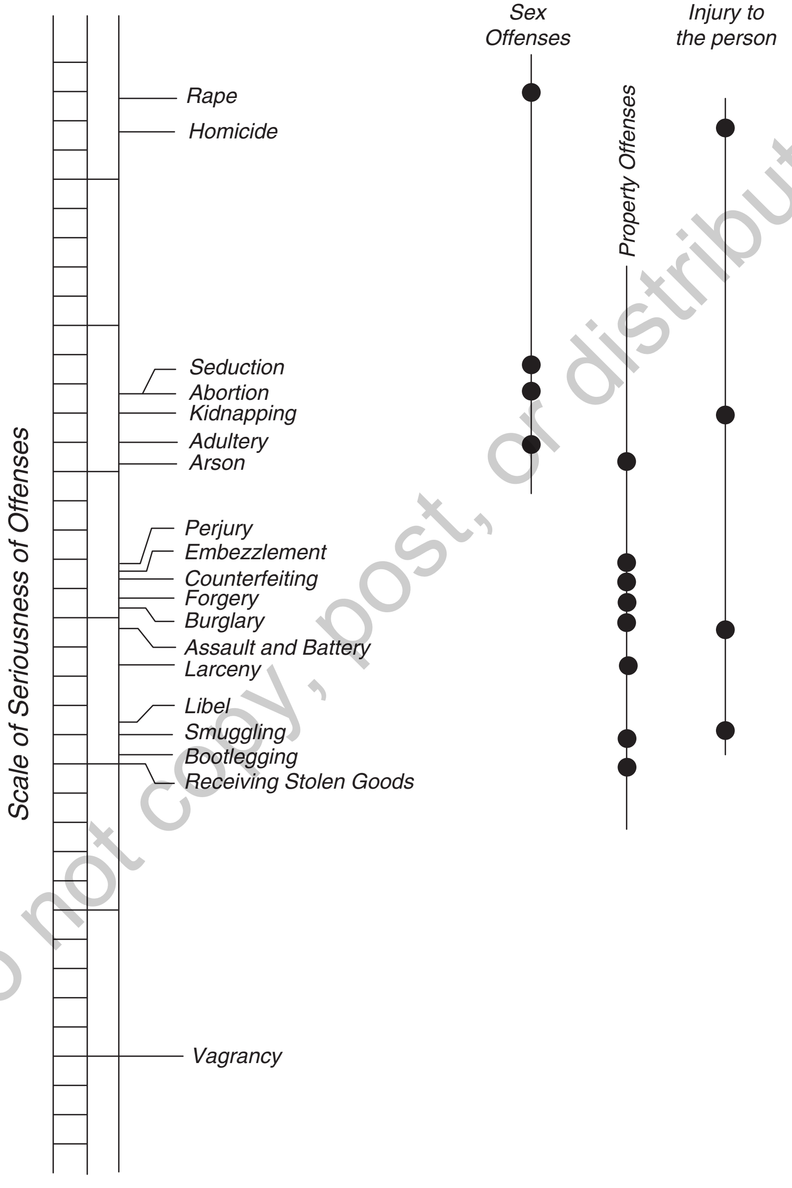
Figure 1.5 presents one of the earliest item maps created by Thurstone (1927) to represent severity of crimes. Several features should be noted. First, Thurstone uses a line as his map for the latent variable of seriousness of offenses. Next, he provides a scaling of the particular items on the latent variable scale. And finally, he provides a conceptual grouping of the items into sex offenses, property offenses, and injury to the person in order to highlight an underlying structure (domains and dimensions for grouping the crimes). As another example, Figure 1.6 provides a Wright map that shows the simultaneous ordering of both items and persons on the latent variable of food insecurity. These data are used throughout this book to illustrate the construction of a scale based on the principles of Rasch measurement theory.

Figure 1.7 presents the components of scale development based on Rasch measurement theory. Central to all of the components is the definition of a latent variable that is manifested in a scale that can be shared and used by a community of scholars. All of the components

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Figure 1.5 Thurstone's Item Map for Seriousness of Criminal Offenses



Source: Thurstone (1927).

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Figure 1.6 Wright Map for Food Insecurity Experience Scale

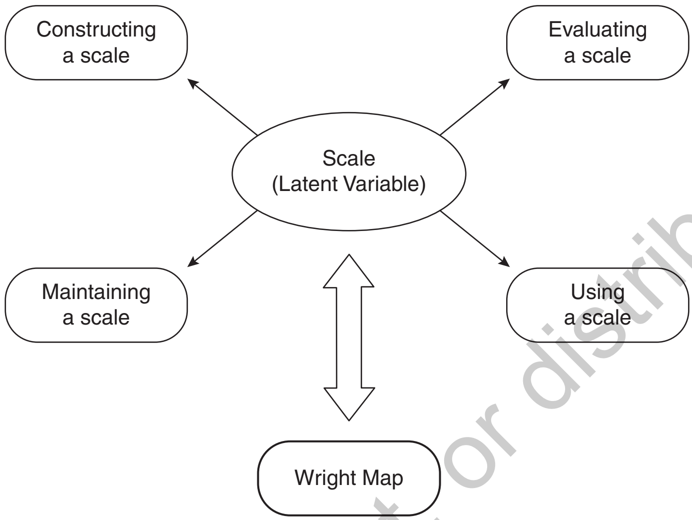
```
|-----------------------------------------------------------------------------------|
|Logit|+Person                                    |-Items                           |
|-----+-------------------------------------------+---------------------------------|
|     | High food insecurity                      | Hard to affirm                  |
|   3 +                                           +                                 |
|     |                                           | Whole Day                       |
|     |                                           |                                 |
|     | 22 23 29                                  |                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|   2 +                                           +                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|     | 13 17 25 33 37                            |                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|   1 +                                           +                                 |
|     |                                           |                                 |
|     |                                           | Hungry                          |
|     | 26 35                                     |                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|     |                                           | Ran Out  Skipped  Worried       |
|     |                                           |                                 |
*   0 * 2 7 9 21                                  *                                 *
|     |                                           |                                 |
|     |                                           |                                 |
|     |                                           | Ate Less                        |
|     |                                           |                                 |
|     |                                           |                                 |
|     | 1 11 24 27 28 36                          |                                 |
|     |                                           |                                 |
|  -1 +                                           +                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|     | 3 4 12 18 20 39                           |                                 |
|     |                                           | Healthy                         |
|     |                                           |                                 |
|     |                                           |                                 |
|  -2 +                                           +                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|     |                                           | Few Foods                       |
|     |                                           |                                 |
|     | 5 6 8 10 14 15 16 19 30 31 32 34 38 40    |                                 |
|     |                                           |                                 |
|     |                                           |                                 |
|  -3 +                                           +                                 |
|     |  Low food insecurity                      | Easy to affirm                  |
|-----------------------------------------------------------------------------------|
```

Note. The full text for the eight items is described in Chapter 2.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Figure 1.7 Components of Scale Development



contribute to the scale that is created to represent the latent variable. Within the context of Rasch measurement theory, the Wright map plays a key role for describing, representing, and defining the latent variable. It should be noted that the components are interconnected, but that each component is likely to be stressed at different times during the scale development process.

The first component in scale development is *constructing the scale*. The guidelines used for this component are inspired by Wilson (2005). We have modified his constructing measures approach in some ways, but the key idea of a construct map (Wright map) is fundamental for this component. We adapt the idea of four building blocks (Wilson, 2005). These four building blocks are the construct map, the item design, the outcome space, and the measurement model. Each of these building blocks is described in detail later in our book with an example of constructing a Food Insecurity Experience (FIE) scale.

Once a scale is constructed, the second component involves *evaluating the scale*. Essentially, this component focuses on empirical examinations of how well the requirements of invariant measurement are met with a particular data set. Rasch measurement theory provides

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

a set of specific requirements that offer the opportunity to achieve item-invariant measurement, person-invariant item calibration, and the creation of a unidimensional scale (Wright map). We use an approach based on the analyses of residuals to critically evaluate model-data fit (Wells & Hambleton, 2016).

The next component is *maintaining the scale*. Researchers seek to develop stable scales that meet the demands of invariant measurement, especially ensuring comparability of measures obtained in different conditions. Once a scale is created and evaluated, it is important to develop a plan and set of principles for maintaining the robustness and comparability of the scores. Some of the main topics include what has been called equating and linking of the scales based on Rasch measurement theory.

The final component is *using the scale*. The reliability, validity, and fairness of the scores obtained from a scale must be examined. Also, it is important to apply the scale in policy contexts that include the setting of performance standards.

It should be stressed that these four components are not always enacted in a strict fashion or in a specified order, and that different components have different levels of emphasis at different stages in scale development. It should also be highlighted that there are interconnections between the various components, and that the components reflect an iterative process of scale development.

1.4 Four Measurement Problems

There are a variety of measurement problems encountered in the social sciences. Many of these problems can be addressed in terms of invariant measurement with Rasch measurement theory as one approach. In this section, we introduce the use of Rasch measurement theory as an aid in conceptualizing four important measurement problems. These measurement problems are developed in greater detail in each chapter of our book. Specifically, we consider the following measurement problems:

- Definition of a latent variable
- Evaluation of differential item functioning
- Examination of interchangeability of items for person measurement
- Creation of performance standards (cut scores) for standard setting

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

One of the basic problems in the social sciences is the selection and definition of latent variables based on substantive theories. How do researchers select their variables? How do researchers select and calibrate various indicators to measure these variables? What guidelines should a community of scholars use to evaluate the quality of the proposed measures? We argue in this book that the use of Rasch measurement theory to create meaningful and useful scales is essential for advancing research in the social, behavioral, and health sciences. The use of arbitrary units (e.g., setting the unit based on an estimate of variance in an arbitrary sample) is unsatisfactory especially when there are clear guidelines for creating scales with meaningful units based on Rasch measurement theory.

Differential item functioning is a measurement problem that violates measurement invariance of a scale (Millsap, 2011). As pointed out by Millsap (2011), "measurement invariance is built on the notion that a measuring device should function the same way across varied conditions, as long as those varied conditions are irrelevant to the attribute being measured" (p. 1). In this book, we view this problem as a failure to meet the requirement of person-invariant calibration of items.

The next measurement problem highlights the interchangeability of items. It is important to recognize the duality of scaling models (Engelhard, 2008a). Measurement invariance is reflected in differential item functioning, but this idea is also relevant for obtaining comparable scores for persons using different subsets of items. The item-invariant person measurement provides the opportunity to achieve measure invariance for persons being independent of specific items. This measurement problem falls under the general category of test equating—the main idea is that the latent variable and the Wright map that represents the scale remain invariant over different sets of observations defined by different items but designed to define the same construct for persons. Computer-adaptive testing is based on the idea that different items are interchangeable, and different subsets of items can be used to obtain comparable measures for each person.

The final measurement problem that we address is standard setting (Cizek, 2012). Standard setting is a process for determining critical points on a scale (cut scores) that represent substantively distinct locations along the latent variable. In many cases, such as educational achievement and food insecurity, ordered categorical reporting of continuous scales is required (fail/pass, food secure/food insecure) to inform policy. Standard setting is the process used to set these levels and create meaningful categories to guide policy makers and other stakeholders.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The chapters of this book provide illustrations of how Rasch measurement theory can be used to solve measurement problems through the components of scale development. Our aim is to describe an approach for problem-solving in measurement emerging from the requirements for invariant measurement and the quest for robust and stable systems of measurement.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.