Behaviormetrics:
Quantitative Approaches to Human Behavior 13

Kojiro Shojima

Test Data Engineering

Latent Rank Analysis, Biclustering, and Bayesian Network

Springer

Behaviormetrics: Quantitative Approaches to Human Behavior

Volume 13

Series Editor

Akinori Okada, Professor Emeritus, Rikkyo University,
Tokyo, Japan

This series covers in their entirety the elements of behaviormetrics, a term that encompasses all quantitative approaches of research to disclose and understand human behavior in the broadest sense. The term includes the concept, theory, model, algorithm, method, and application of quantitative approaches from theoretical or conceptual studies to empirical or practical application studies to comprehend human behavior. The Behaviormetrics series deals with a wide range of topics of data analysis and of developing new models, algorithms, and methods to analyze these data.

The characteristics featured in the series have four aspects. The first is the variety of the methods utilized in data analysis and a newly developed method that includes not only standard or general statistical methods or psychometric methods traditionally used in data analysis, but also includes cluster analysis, multidimensional scaling, machine learning, corresponding analysis, biplot, network analysis and graph theory, conjoint measurement, biclustering, visualization, and data and web mining. The second aspect is the variety of types of data including ranking, categorical, preference, functional, angle, contextual, nominal, multi-mode multi-way, contextual, continuous, discrete, high-dimensional, and sparse data. The third comprises the varied procedures by which the data are collected: by survey, experiment, sensor devices, and purchase records, and other means. The fourth aspect of the Behaviormetrics series is the diversity of fields from which the data are derived, including marketing and consumer behavior, sociology, psychology, education, archaeology, medicine, economics, political and policy science, cognitive science, public administration, pharmacy, engineering, urban planning, agriculture and forestry science, and brain science.

In essence, the purpose of this series is to describe the new horizons opening up in behaviormetrics — approaches to understanding and disclosing human behaviors both in the analyses of diverse data by a wide range of methods and in the development of new methods to analyze these data.

More information about this series at https://link.springer.com/bookseries/16001

Kojiro Shojima

Test Data Engineering

Latent Rank Analysis, Biclustering, and Bayesian Network

Springer

Kojiro Shojima
Research Division
National Center for University Entrance Examinations
Tokyo, Japan

This work was supported by JSPS KAKENHI Grant Number JP20K03383

ISSN 2524-4027 ISSN 2524-4035 (electronic)
Behaviormetrics: Quantitative Approaches to Human Behavior
ISBN 978-981-16-9985-6 ISBN 978-981-16-9986-3 (eBook)
https://doi.org/10.1007/978-981-16-9986-3

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

To Sachiko, Sāya, and Yūka

Contents

Abbreviations, Symbols, and Notations

Abbreviations

1PLM	1-parameter logistic model
2PLM	2-parameter logistic model
3PLM	3-parameter logistic model
4PLM	4-parameter logistic model
AIC	Akaike information criterion
AM	Analysis model
AMDS	Asymmetric multidimensional scaling
BIC	Bayesian information criterion
BINET	Bicluster network model
BM	Benchmark model
BNM	Bayesian network model
BRM	Bicluster reference matrix
BVN	Bivariate normal
CAIC	Consistent AIC
CAT	Computer adaptive test
CBT	Computer-based test
CCRR	Conditional correct response rate
CRR	Correct response rate
CDF	Cumulative distribution function
CDM	Cognitive diagnostic model
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CIRR	Conditional incorrect response rate
CMD	Class membership distribution
CP	Conditional probability
CRM	Class reference matrix
CRP	Chinese restaurant process
CRR	Correct response rate

CRV	Class reference vector
CTT	Classical test theory
DAG	Directed acyclic graph
DIF	Differential item functioning
DF	Degrees of freedom
EAP	Expected a posteriori
EDF	Effective degrees of freedom
EFA	Exploratory factor analysis
ELL	Expected log-likelihood
ELP	Expected log-posterior
EM algorithm	Expectation-maximization algorithm
FRP	Field reference profile
GA	Genetic algorithm
GTM	Generative topographic mapping
ICC	Item characteristic curve (=IRF)
IFI	Incremental fit index
IIF	Item information function
IQR	Interquartile range
IRC	Item-remainder correlation
IRF	Item response function
IRM	Infinite relational model
IRP	Item reference profile
IRR	Incorrect response rate
IRT	Item response theory
ITC	Item-total correlation
JCRR	Joint correct response rate
JP	Joint probability
KR-20	Kuder-Richardson formula 20 (reliability coefficient)
LCA	Latent class analysis
LCD	Latent class distribution
LDB	Local dependence biclustering
LD-LCA	Local dependence LCA
LD-LRA	Local dependence LRA
LDR	Local dependence ranklustering
LFA	Latent field analysis
LFD	Latent field distribution
LHS	Left-hand side
LRA	Latent rank analysis
LRD	Latent rank distribution
LRE	Latent rank estimate
MAP	Maximum a posteriori
MIR	Monotonic IRP ratio
ML	Maximum likelihood
MLE	Maximum likelihood estimate (or estimation)
MVN	Multivariate normal

NaN	Not a number
NFI	Normed fit index
NM	Null model
NP	Number of parameters
NRS	Number-right score
PBIL	Population-based incremental learning
PDF	Probability density function
PIRP	Parent item response pattern
POS data	Point of sales data
PPT	Paper-and-pencil test
PSD	Posterior standard deviation
PSR	Passing student rate
PSRP	Parent student response pattern
RDO	Rank-down odds
RFI	Relative fit index
RHS	Right-hand side
RMD	Rank membership distribution
RMP	Rank membership profile
RMSEA	Root mean square error of approximation
RRM	Rank reference matrix
RRV	Rank reference vector
RUO	Rank-up odds
SD	Standard deviation
SE	Standard error
SEM	Structural equation modeling
SOAC	Strongly ordinal alignment condition
SOM	Self-organizing map
TCC	Test characteristic curve (=TRF)
TDA	Test data analysis
TDE	Test data engineering
TF	True or false
TIF	Test information function
TLI	Tucker-Lewis index
TPS	Total passing students
TRF	Test response function
TRP	Test reference profile
TVN	Trivariate normal
WOAC	Weakly ordinal alignment condition

Symbols

$\wedge$	And (logical conjunction)
$\vee$	Or (logical disjunction)

$\sim$	$X \sim D$ implies "X is randomly sampled from distribution D"
$\approx$	Approximately equal to
$\propto$	Proportional to
$:=$	Defined as: e.g., $LHS := RHS$ implies "LHS is defined as RHS"
$\Leftrightarrow$	If and only if: $A \Leftrightarrow B$ implies "A is true (false) if and only if B is true (false)"
$\gtrless$	For real numbers a, b, c, and d, $a \gtrless b \Leftrightarrow c \gtrless d$ implies "$a > b$ if $c > d$ and $a < b$ if $c < d$"
$\therefore$	Therefore
$\because$	Because
$\lvert\cdot\rvert$	Absolute value: for a real number a, $\lvert a\rvert$ is the non-negative value of a, e.g., $\lvert-3.14\rvert = 3.14$ and $\lvert 2.72\rvert = 2.72$ (unchanged when a is non-negative)
$\lvert\lvert\cdot\rvert\rvert$	Norm: in this book, it denotes a vector's Euclid distance from the origin; for a vector $\boldsymbol{a} = [a_1 \cdots a_J]'$, $\lvert\lvert\boldsymbol{a}\rvert\rvert = \sqrt{\sum_{j=1}^{J} a_j^2}$
$\lceil\cdot\rceil$	Ceiling function: rounds the argument to the smallest integer but not smaller than the argument, e.g., $\lceil 3.14\rceil = 4$.
$\lfloor\cdot\rfloor$	Floor function: rounds the argument to the largest integer but not larger than the argument, e.g., $\lfloor 3.14\rfloor = 3$.
round$(\cdot)$	Rounding (half away from 0) function: rounds the argument to the nearest integer, e.g., round$(3.4) = 3$, round$(3.5) = 4$, and round$(-3.5) = -4$
Π	Repeated multiplication: for a series of numbers $\{a_1, a_2, \cdots, a_N\}$, $\prod_{n=3}^{6} a_n = a_3 \times a_4 \times a_5 \times a_6$
!	Factorial: $n! = \prod_{i=1}^{n} i = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$
$\odot$	Hadamard product: conducts element-wise product of two matrices of the same size; if $\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$, then $\boldsymbol{A} \odot \boldsymbol{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & a_{13}b_{13} \\ a_{21}b_{21} & a_{22}b_{22} & a_{23}b_{23} \end{bmatrix}$
$\oslash$	Hadamard division: conducts element-wise division of two matrices of the same size; if $\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$, then $\boldsymbol{A} \oslash \boldsymbol{B} = \begin{bmatrix} a_{11}/b_{11} & a_{12}/b_{12} \\ a_{21}/b_{21} & a_{22}/b_{22} \end{bmatrix}$
$\circ$	Hadamard power: raises each element of a matrix to a power; for a matrix $\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$ and real number b, $\boldsymbol{A}^{\circ b} = \begin{bmatrix} a_{11}^b & a_{12}^b & a_{13}^b \\ a_{21}^b & a_{22}^b & a_{23}^b \end{bmatrix}$, when $b = \frac{1}{2}$, it is particularly referred to as Hadamard root.
$\forall$	For all: $\forall j \in \mathbb{N}_J$ implies "for all numbers in the set of natural numbers $\mathbb{N}_J = \{1, 2, \cdots, J\}$," which implies "for all items" in this book; similarly, $\forall s \in \mathbb{N}_S$ represents for "for all students."

$\in$	In or member (element) of: $2 \in \mathbb{N}$ implies "2 is a member of natural numbers"; $0.3 \in [0, 1]$ represents "0.3 is in (the range of) 0 to 1"
$\backslash$	Set difference or relative complement: if $A = \{a_1, a_2, a_3, a_4\}$ and $B = \{a_1, a_3\}$, then $A \backslash B = \{a_1, a_3\}$
arg max	$\hat{x} = \arg\max_{x \in \mathbb{R}} f(x)$ implies $\hat{x}$ is the value in $\mathbb{R}$ that maximizes f
$\mathrm{cl}(\cdot)$	Clip function or clamp function: $\mathrm{cl}(x) = x$ when $x_{min} \le x \le x_{max}$, $\mathrm{cl}(x) = x_{min}$ when $x < x_{min}$, and $\mathrm{cl}(x) = x_{max}$ when $x > x_{max}$, where $(x_{min}, x_{max}) = (0, 1)$ in this book
$\det(\cdot)$	Determinant: $\det(\boldsymbol{A})$ denotes the determinant of matrix $\boldsymbol{A}$
$\mathrm{diag}(\cdot)$	Extracts the diagonal elements of a square matrix and vectorizes them: for $\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, $\mathrm{diag}(\boldsymbol{A}) = \begin{bmatrix} a_{11} \\ a_{22} \end{bmatrix}$.
$E[\cdot]$	Expectation: the expectation or the expected value of a random variable X, denoted $E[X]$, is intuitively the arithmetic mean of a large number of independently sampled relalizations of X; the expectation for a discrete random variable is the probability-weighted sum of all possible outcomes
$\exp(\cdot)$	Exponential: for a real number a, $exp(a)$ is the same as e^a, where $e = 2.71828\cdots$ is Napier's constant; $exp(3) = 2.718^3 = 20.09$
$\ln(\cdot)$	Natural logarithm: the logarithm with base $e = 2.71828\cdots$; $\ln a = \log_e a$
$\mathrm{row}(\cdot)$	Returns the number of rows of the argument (matrix): e.g., $\mathrm{row}(\boldsymbol{A}) = 3$ when $\boldsymbol{A}$ is a 3×2 matrix; returns the number of elements when the argument is a vector
$\mathrm{sgn}(\cdot)$	Signum (or sign) function: returns the argument's sign, i.e., $\mathrm{sgn}(x) = 1$ if $x > 0$, $\mathrm{sgn}(x) = 0$ if $x = 0$, and $\mathrm{sgn}(x) = -1$ if $x < 0$
$\mathrm{tr}(\cdot)$	Trace: the sum of the argument (a square matrix); when $\boldsymbol{A}$ is a $J \times J$ matrix, $\mathrm{tr}(\boldsymbol{A}) = \sum_{j=1}^{n} a_{jj}$
$\mathbb{N}$	The set of all natural numbers, $\{1, 2, 3, \cdots\}$: the set with subscript n ($\mathbb{N}_n$) represents the set of natural numbers from 1 to n, $\{1, 2, \cdots, n\}$
$\emptyset$	The empty set: the set has no elements, and the size of this set is zero

Notations

$\mathbf{1}_n$	A vector of size n, in which all elements are 1:$\mathbf{1}_3 = [1\ 1\ 1]'$
$\boldsymbol{A} = \{a_{jk}\}$	Adjacency matrix, where $a_{jk} = 1$ when an edge of Item $j \to k$ exists; otherwise $a_{jk} = 0$
a_j	Slope parameter of Item j (IRT)
$\tilde{\alpha}_j$	Slope index of Item j (LRA): the lower rank of two adjacent ranks between which the CRR increases the most in the IRP of Item j
$\tilde{a}_j$	CRR difference between Ranks α_j and $\alpha_j + 1$ (LRA)
b_j	Location parameter of Item j (IRT)

$\tilde{\beta}_j$	Location index of Item j (LRA): the rank whose CRR is the closest to 0.5
$\tilde{b}_j$	CRR at Rank β_j in IRP of Item j (LRA)
C	Number of latent classes (LCA)
c_j	Lower asymptote parameter of Item j (IRT)
$\widetilde{\gamma}_j$	Monotonicity index of Item j (LRA): the ratio of adjacent rank pairs between which CRR decreases to the number of adjacent rank pairs
$\tilde{c}_j$	Cumulative decrease of adjacent rank pairs between which CRR decreases (LRA)
d_j	Upper asymptote parameter of Item j
$\boldsymbol{\Delta}$	A diagonal matrix where eigenvalues are arranged in the descending order
δ_j	j-th largest eigenvalue or j-th diagonal element in $\boldsymbol{\Delta}$
$\boldsymbol{E} = \{\boldsymbol{e}_j\}$	A matrix in which the j-th row vector is $\boldsymbol{e}_j$, which is the eigenvector corresponding to δ_j
$f_N(\cdot;\mu,\sigma)$	PDF of normal distribution with mean μ and SD σ (variance σ^2); $f_N(\cdot;0,1)$ is the PDF of the standard normal distribution
$F_N(\cdot;\mu,\sigma)$	CDF of normal distribution with mean μ and SD σ; $F_N(\cdot;0,1)$ is the CDF of the standard normal distribution
$F_N^{-1}(\cdot;\mu,\sigma)$	Quantile function or inverse CDF of normal distribution with mean μ and SD σ; $F_N^{-1}(\cdot;0,1)$ is the quantile function of the standard normal distribution
$f_{N^2}(\cdot,\cdot;\rho)$	PDF of standard bivariate normal distribution with correlation ρ
$\boldsymbol{\Gamma} = \{\gamma_{jfcd}\}$	Four-dimensional PSRP array (BINET), where γ_{jfcd} is 1 when Item j is the d-th item in Field f, and Class c is locally dependent at the field; otherwise γ_{jfcd} is 0
$\boldsymbol{\Gamma} = \{\gamma_{sjd}\}$	Three-dimensional PIRP array (BNM), where γ_{sjd} is 1 when the response pattern of Student s for Item j's parent item(s) is the d-th pattern; otherwise, γ_{sjd} is 0
$\boldsymbol{\Gamma} = \{\gamma_{srjd}\}$	Four-dimensional PIRP array (LD-LRA), where γ_{srjd} is 1 when the response pattern of Student s in Rank r for Item j's parent item(s) is the d-th pattern; otherwise, γ_{srjd} is 0
$\boldsymbol{\Gamma} = \{\gamma_{srfd}\}$	Four-dimensional PIRP array (LDB), where γ_{srfd} is 1 when the NRS of Student s in Rank r for the items classified in Field f's parent field(s) is the d-th pattern; otherwise, γ_{srfd} is 0
$\boldsymbol{I}_n$	Identity matrix of dimensions $n \times n$
$\boldsymbol{I}_j(\boldsymbol{\lambda})$	Posterior information matrix of Item j, where $\boldsymbol{\lambda}$ is an item parameter vector (IRT)
$\boldsymbol{I}_j^{(F)}(\boldsymbol{\lambda})$	Fisher information matrix of Item j, where $\boldsymbol{\lambda}$ is an item parameter vector (IRT)
$\boldsymbol{I}^{(pr)}(\boldsymbol{\lambda})$	Prior information matrix, where $\boldsymbol{\lambda}$ is an item parameter vector (IRT)
J	Number of items (test length)
$\boldsymbol{\lambda}_j$	Item parameter vector of Item j (IRT)
$\boldsymbol{\Lambda}$	Item parameter matrix containing $\boldsymbol{\lambda}_j$ in the j-th row (IRT)

$\boldsymbol{M}_C = \{m_{sc}\}$	Class membership matrix, where the (s, c)-th element, m_{sc}, represents the membership (probability) that Student s belongs to Class c (LCA)
$\boldsymbol{M}_F = \{m_{jf}\}$	Field membership matrix, where m_{jf} represents the membership (probability) that Item j is classified in Field f (biclustering)
$\boldsymbol{M}_G = \{m_{sg}\}$	Group membership matrix, where m_{sg} is 1 if Student s belongs to Group g; otherwise, m_{sg} is 0
$\boldsymbol{M}_R = \{m_{sr}\}$	Rank membership matrix, where m_{sr} represents the membership (probability) that Student s belongs to Rank r (LRA)
$\boldsymbol{o} = \{o_j\}$	Item odds vector, in which o_j is the odds of Item j
$\boldsymbol{p} = \{p_j\}$	CRR vector, in which p_j is the CRR of Item j
$\boldsymbol{p}^{(w)} = \{p_j^{(w)}\}$	Item mean vector, in which $p_j^{(w)}$ is the mean of Item j
$\boldsymbol{P}_C = \{p_{k\vert j}\}$	CCRR matrix: an asymmetric matrix $(\boldsymbol{P}'_C \neq \boldsymbol{P}_C)$ with all diagonal elements being 1, and the (j, k)-th entry, $p_{k\vert j}$, is the CRR of the students passing Item k for Items j
$\boldsymbol{P}_G = \{p_{jg}\}$	Group reference matrix, where p_{jg} is the CRR of students belonging to Group g for Item j
$\boldsymbol{P}_J = \{p_{\overline{jk}}\}$	JCRR matrix: a symmetric matrix $(\boldsymbol{P}'_J \neq \boldsymbol{P}_J)$, where the (j, k)-th element, $p_{\overline{jk}}$, is the JCRR of Items j and k, and the j-th diagonal element is the CRR of Item j
$\boldsymbol{P}_L = \{l_{jk}\}$	Item lift matrix: a symmetric matrix $(\boldsymbol{P}'_L = \boldsymbol{P}_L)$, where l_{jk} is the lift of Item $j \to k$
$P(\theta; \boldsymbol{\lambda}_j)$	IRF of item j with item parameter $\boldsymbol{\lambda}_j$ (IRT), also denoted as $P_j(\theta)$
$Q(\theta; \boldsymbol{\lambda}_j)$	Incorrect response function of Item j identical to $1-P(\theta; \boldsymbol{\lambda}_j)$ (IRT), also denoted as $Q_j(\theta)$
$\boldsymbol{\Phi} = \{\phi_{jk}\}$	A symmetric matrix $(\boldsymbol{\Phi}' = \boldsymbol{\Phi})$ with diagonal elements 1, where ϕ_{jk} is the ϕ coefficient between Items j and k
$\boldsymbol{\Pi}_B = \{\pi_{cf}\}$	Bicluster reference matrix (biclustering), where π_{cf} is the CRR of Class c students for a Field f item
$\boldsymbol{\Pi}_C = \{\pi_{cj}\}$	Class reference matrix (LCA), where π_{cj} is the CRR of Class c students for Item j
$\boldsymbol{\Pi}_{DC} = \{\pi_{fcj}\}$	Local dependence parameter set (BINET), where π_{fcj} is the PSR of Class c students for Item j in Field f
$\boldsymbol{\Pi}_{DF} = \{\pi_{rfd}\}$	Local dependence parameter set (LDB), where π_{rfd} is the CRR for a Field f item of Rank r students whose PIRP is the d-th pattern
$\boldsymbol{\Pi}_{DI} = \{\pi_{rjd}\}$	Local dependence parameter set (LD-LRA), where π_{rjd} is the CRR for Item j of Rank r students whose PIRP is the d-th pattern
$\boldsymbol{\Pi}_\Gamma = \{\pi_{jd}\}$	Parameter set (BNM), where π_{jd} is the CRR for Item j of students whose PIRP is the d-th pattern
$\boldsymbol{\Pi}_R = \{\pi_{jr}\}$	Rank reference matrix (LRA), where π_{jr} is the CRR of Rank r students for Item j
R	Number of latent ranks (LRA)
$\boldsymbol{R}$	(Tetrachoric) correlation matrix

$\boldsymbol{r} = \{r_s\}$	Student passage rate vector, where r_s is the passage rate of Student s for test items
$\boldsymbol{r}^{(w)} = \left\{r_s^{(w)}\right\}$	Student scoring rate vector, where $r_s^{(w)}$ is the scoring rate of Student s
$\boldsymbol{\rho}_\zeta = \left\{\rho_{j,\zeta}\right\}$	Item-total correlation vector, where the j-th element, $\rho_{j,\zeta}$, is the item-total correlation of $\boldsymbol{u}_j$ with $\boldsymbol{\zeta}$
S	Sample size (number of students)
$\boldsymbol{t} = \{t_s\}$	Student NRS vector, where t_s is the NRS of Student s
$\boldsymbol{t}^{(w)} = \left\{t_s^{(w)}\right\}$	Student total score vector, where $t_s^{(w)}$ is the total score of Student s
$\boldsymbol{t}_{TRP}$	TRP, which is a column sum vector of class/rank/bicluster reference matrix and represents the expected NRS on the test
$\boldsymbol{t}_{TRP}^{(w)}$	Weighted TRP, representing the expected score on the test
$\boldsymbol{\tau} = \{\tau_s\}$	Student true score vector, where τ_s is the true score of Student s
$\boldsymbol{\theta} = \{\theta_s\}$	Student ability value vector, where θ_s is the ability value of Student s (IRT)
$\boldsymbol{U} = \left\{u_{sj}\right\}$	Data matrix, where u_{sj} is the response of Student s to Item j, coded 1 if it is correct, 0 if incorrect, and "(dot)" if missing (the item is not presented to the student)
$\boldsymbol{U}_\Gamma = \left\{u_{\Gamma sj}\right\}$	PIRP matrix, where the (s, j)-th entry, $u_{\Gamma sj}$, is d when Student s's PIRP of Item j is the d-th pattern
$\boldsymbol{u}_j$	Data vector of Item j or the j-th column vector in $\boldsymbol{U}$
$\boldsymbol{u}_s$	(Vertical) vector of Student s's data arranged in the s-th row of $\boldsymbol{U}$
$V^{(w)}$	Variance of total scores, or $Var\left[\boldsymbol{t}^{(w)}\right]$
$V^{(\tau)}$	Variance of true scores, or $Var[\boldsymbol{\tau}]$
$V^{(e)}$	Error variance, or $Var[\boldsymbol{e}]$
$\boldsymbol{v} = \left\{v_j\right\}$	Item variance vector, where v_j is the variance of Item j
$\boldsymbol{w} = \left\{w_j\right\}$	Item weight vector, where w_j is the weight of Item j
$\boldsymbol{Z} = \left\{z_{sj}\right\}$	Missing indicator matrix, where z_{sj} is 1 if Item j is presented to Student s; otherwise, z_{sj} is 0
$\boldsymbol{z}_j$	Missing indicator vector for Item j or the j-th column vector of $\boldsymbol{Z}$
$\boldsymbol{z}_s$	(Vertical) vector of Student s's missing indicators arranged in the s-th row of $\boldsymbol{Z}$
$\boldsymbol{\zeta} = \{\zeta_s\}$	Vector of standardized scores of students' passage rates, where ζ_s is the standardized score of Student s's passage rate
$\boldsymbol{\zeta}^{(w)} = \left\{\zeta_s^{(w)}\right\}$	Vector of standardized scores of students' scoring rates, where $\zeta_s^{(w)}$ is the standardized score of Student s's scoring rate

Chapter 1
Concept of Test Data Engineering

When managing a test, item-writing (or question-making) and scoring (or marking) are the main segments; however, there is a whole package of design, including temporal planning and spatial layout, item-writing, implementation, scoring, analysis, evaluation, and feedback. Among them, the term "test data analysis" is an expression that focuses on the analysis component; however, "test data engineering," a concept considered in this book, represents a broader perspective.

Simply put, tests are tools used in society. They are highly public in nature; thus, it can be said to be public tools. They sometimes affect the future of test takers, most of whom are children and young people. Thus, they should neither be groundless nor irresponsible, and adults must properly summarize the information contained in the test data and provide them back to test takers. In this book, test takers are referred to as "students."

The primary information obtained from a test administration is the raw data; i.e., students' responses. If an item is "What is the capital of USA?," the responses are the uncoded raw responses of the students, such as "New York," "Washington, D.C.," and "London." The responses include "missing data" and "nonresponses," which are also important information about the students. Secondary information is marked (coded) data. Except for essay questions,[1] each test datum is trichotomous: correct, incorrect, or missing. The correct and incorrect responses are normally coded as 1 and 0, respectively, while the missing datum is denoted by various symbols such as 99, -1, and "." (dot). In addition, a nonresponse is usually regarded as an incorrect response and coded as 0 (see also Table 2.2, p. 16). This book mainly focuses on analyzing secondary information (post-coded data).

[1] This book does not deal with essay questions. An essay question is often graded from points of view. For example, if an essay item has five viewpoints, and each viewpoint is assigned three points, the scale of the essay item is 0 to 15.

K. Shojima, *Test Data Engineering*, Behaviormetrics: Quantitative Approaches to Human Behavior 13, https://doi.org/10.1007/978-981-16-9986-3_1

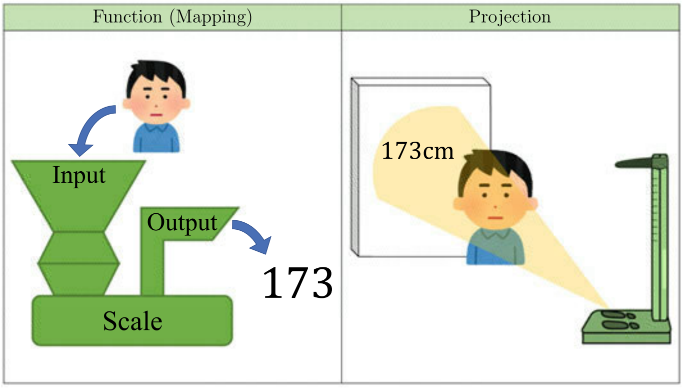

Fig. 1.1 Analogy of measurement

Generally, changing the state of something by retouching it is called “processing”; therefore, changing the primary information (raw data) to secondary information (coded data) is also considered as a process. In addition, counting the number of correct responses for each student and calculating the correct response rate for each item are also processes, and they may be referred to as the “tertiary information briefing the secondary information” and are often used as the summary and feedback returned to students and the teacher who administered the test. In other words, a test administration is considered as a series of processes pertaining to collecting, coding, scoring, and transforming data.

1.1 Measurement as Projection

Generally, measurement is the action of assigning a numerical value to an object.[2] For example, suppose that Mr. Smith’s height is measured and found to be 173 cm (5′8″), then this can be formulated as

$$f_{\text{SCALE}}(\text{MR. SMITH}) = 173\ (\text{cm}). \tag{1.1}$$

This indicates that the height scale can be regarded as a function that outputs “173 (cm)” for the input “Mr. Smith.” The function f_{SCALE} can also be considered as a machine that extracts only his body length (Fig. 1.1, left) but discards all other information, such as gender, age, nationality, ethnicity, birthday, address, and hobbies.

[2] For a historical background on educational measurement theory, see (Clauser and Bunch, 2022). This chapter describes the author’s thought.

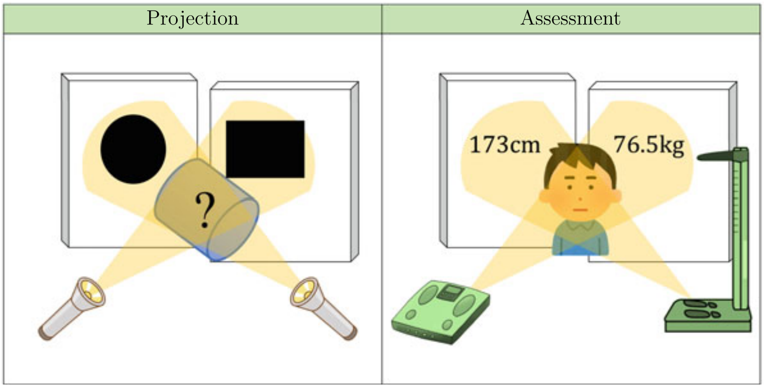

Fig. 1.2 Measurement as projection

Equation (1.1) can alternatively be represented as

$$\text{MR. SMITH} \xmapsto{f_{\text{SCALE}}} 173 \text{ (cm)}.$$

The two formulations are different representations conveying the same meaning, where $A \overset{f}{\mapsto} B$ means that A is transformed (or **mapped** in a technical term) to B by f. For a more literal sense, it can be read as "Painter f studied Scenery A and drew Picture B," or "when A was illuminated from the direction of f, the shadow (image) of B appeared in the back." The latter expression is technically referred to as a **projection** (Fig. 1.1, right).

As Fig. 1.2 (left) illustrates, illuminating an object from various angles, examining the shadows (flattened images on the respective walls) of the object, and considering the original shape of the object is referred to as **projection pursuit**. In particular, when the shape of an object is complex and multidimensional, the object cannot be observed directly; in this case, the original shape can be imagined by inspecting shadow images cast by light from various directions.

The measurement is considered a projection. "Mr. Smith" is essentially a high-dimensional input composed of various pieces of information,[3] Thus, it is necessary to illuminate (measure) Mr. Smith from various angles to understand him (see Fig. 1.2, right). When Mr. Smith is illuminated by the projector (from the direction) at "height meter," then image "173 cm" is cast on the wall; and if he is illuminated by "weight scale," then "76.5 kg" is cast on the wall. Thus, we can depict the true shape of Mr. Smith by illuminating him with more projectors. A set of measurements (i.e., projectors) is referred to as an "assessment." In a medical assessment, the doctor

[3] It contains biological (gender, age, height, weight, eyesight, skin color, grip strength, voice pitch), socioeconomic (address, educational history, income, number of siblings, marital status), psychological (agreeableness, conscientiousness, extraversion, neuroticism, openness), and intellectual (language, mathematics, science, history) characteristics.

obtains the shadows of Mr. Smith's health status, an extremely high-dimensional object, using multiple projectors such as blood tests, urinalysis, radiography, and echocardiography, and then integrates the shadows and diagnoses the status.

1.2 Testing as Projection

Likewise, a test can likewise be viewed as a function or projection. Suppose Mr. Smith took a test and scored 73; this situation can be represented as follows:

$$f_{\text{TEST}}(\text{MR. SMITH}) = 73 \text{ (points)}. \tag{1.2}$$

This test can also be regarded as a projector illuminating Mr. Smith by casting the image "73." If this test is a math test, the number expresses the mathematical ability. Moreover, some tests may return a grade such as A, B, or C. Note that f_{TEST} only extracts the object of interest (i.e., mathematical ability) and discards (or neglects) all other information.

Let us express Eq. (1.2) in finer detail. Because the test consists of some items (e.g., 100 items), a correct/incorrect (1/0) response pattern for the 100 items is first output by f_{TEST}, as follows:

$$f_{\text{TEST}}(\text{MR. SMITH}) = \overbrace{11001011011 \cdots 1}^{\text{100 Items}}.$$

This example shows that Mr. Smith passed 73 items (e.g., Items 1, 2, 5, 7, 8, $\cdots$), and failed the other 27 items. Then, by inputting the response pattern into function "number-right score," f_{NRS}, the function finally returns "73" as follows:

$$f_{\text{NRS}}(11001011011 \cdots 1) = 73 \text{ (points)}.$$

This f_{NRS} is also regarded as a projector because a different image is cast by a different projector. For example, if one uses the projector "correct response rate," f_{CRR}, the response pattern is transformed into

$$f_{\text{CRR}}(11001011011 \cdots 1) = 0.73.$$

In addition, suppose he did not respond to 12 of the 100 items. By the projector "nonresponse rate," f_{NRR}, the response pattern is processed into

$$f_{\text{NRR}}(11001011011 \cdots 1) = 0.12.$$

Both 0.73 and 0.12 are the shadows (images) cast by the projectors, respectively; more specifically, f_{TEST} is classified in measurement, while f_{NRS}, f_{CRR}, and f_{NRR}

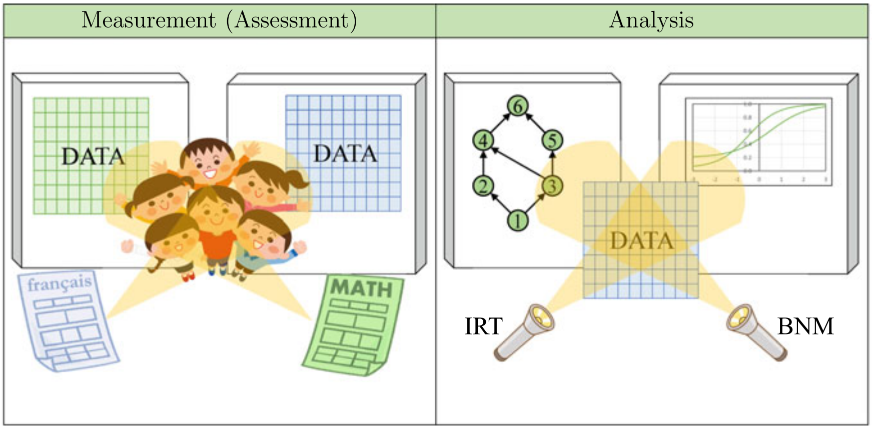

Fig. 1.3 Projection by measurement and analysis

are grouped in the analysis. Accordingly, the process of administering a test to obtain student scores includes basically two projections: measurement and analysis.

Accordingly, Eq. (1.2) is decomposed as follows:

$$f_{\text{NRS}}(f_{\text{TEST}}(\text{MR. SMITH})) = f_{\text{NRS}}(11001011011\cdots 1) = 73 \text{ (points)},$$
$$\text{MR. SMITH} \xmapsto{f_{\text{TEST}}} (11001011011\cdots 1) \xmapsto{f_{\text{NRS}}} 73 \text{ (points)}.$$

These two equations are identical; i.e., they represent the object (i.e., Mr. Smith) that is first measured as (transformed into, output as, symbolized as, coded to, or induced to) the binary response pattern through the test, and the pattern is then analyzed as (processed into, or calculated as) 73.

Usually, a test is performed by some students (Fig. 1.3, left), which is expressed as follows:

$$f_{\text{TEST}}(\text{STUDENTS}) = \text{DATA}, \quad \text{or} \quad \text{STUDENTS} \xmapsto{f_{TEST}} \text{DATA}.$$

In these equations, DATA denotes a data matrix (Sect. 2.1, p. 15), in which Student s's response to Item j is placed in the s-th row and j-th column. The figure also shows that a different aspect (i.e., response pattern) of the students is projected if the same students took a different test, which can be represented as follows:

$$f_{\text{MATH}}(\text{STUDENTS}) = \text{DATA}_M, \text{ or STUDENTS} \xmapsto{f_{MATH}} \text{DATA}_M, \quad (1.3)$$
$$f_{FRENCH}(\text{STUDENTS}) = \text{DATA}_F, \text{ or STUDENTS} \xmapsto{f_{FRENCH}} \text{DATA}_F.$$

Furthermore, Fig. 1.3 (right) shows that different images (i.e., shadows) are generated even for the same data when different analyses are conducted. In the figure, the item response theory (IRT; Chap. 4, p. 85) and Bayesian network model (BNM; Chap. 8, p. 85) are employed. This situation is represented as follows:

$$f_{\text{IRT}}(\text{DATA}) = \text{IRF}, \quad \text{or} \quad \text{DATA} \xmapsto{f_{\text{IRT}}} \text{IRF},$$
$$f_{\text{BNM}}(\text{DATA}) = \text{DAG}, \quad \text{or} \quad \text{DATA} \xmapsto{f_{\text{BNM}}} \text{DAG}.$$

When data are analyzed by IRT, one of the outputs is the item response function (IRF; Sect. 4.2, p. 86).[4] In addition, the BNM yields an output known as directed acyclic graph (DAG; Sect. 8.1.8, p. 366).

A single test administration is a single-facet assessment, while a multifacet assessment is used to measure each student via multiple tests. In general, teachers should inspect each student from multiple angles by employing various evaluation methods. For instance, for an entrance exam, practical skills, interviews, and activity history should be included and integrated into paper-and-pencil tests (PPTs) to determine the treatment (pass or fail) of each applicant. Although each shadow (i.e., image) is a plane, the original shape of the student, as an essentially high-dimensional being, can be approximated from multiple planes.

Design Concept

It seems a matter of course that different analyses project different images, this is because each analysis method was developed based on a unique design concept. The design concept of a method is its policy, philosophy, or specification of the information to be extracted from the data, based on the purpose of the analysis. It also consists of some detailed assumptions (or constraints). For example, the typical assumptions used in IRT are unidimensionality (Sect. 2.8, p. 59) and local independence (Sect. 4.3.1, p. 96). A representative assumption for BNM is d-separation (Sect. 8.2, p. 369).

Once a statistical analysis is conducted, some outputs are produced; thus, a large quantity of information can be extracted from the data by the method. However, it should be noted that what one method can extract from data is only a small part of the entire information contained in the data, and a large part of it is discarded (or neglected), because data contain a vast amount of information. For example, in IRT, information about the local dependency structure that generally exists among items is neglected and abandoned under the assumption of local independence; the local dependency information is shaped (transformed or processed) into a piece of information as if items are locally independent, and such artificial results are finally output.

[4] Various other results can also be output.

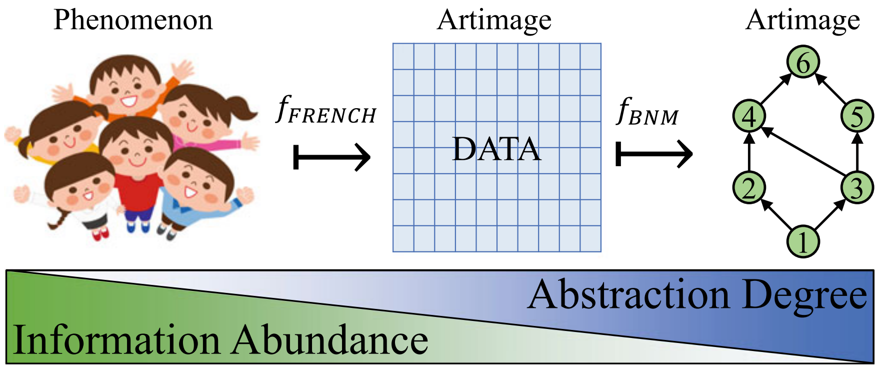

Fig. 1.4 Artimage morphing

1.3 Artimage Morphing

Measurement and analysis are the processes of retouching the input information and remakeing it into a different product. Specifically, measurement processes an object (phenomenon or event) into an image represented by numbers (i.e., data), and analysis processes the data into images of statistical outputs. In particular, testing, which is a measurement method, transforms the input "students" into the image "numerical array" (i.e., binary true/false data); test data analysis then processes the input "binary data" into images such as "indices," "figures," "tables," "information," "knowledge," and "words (including numbers)." Note that all images obtained through such processes are not natural but artificial; thus, they are called **artimages** Shojima 2007a.[5]

Figure 1.4 depicts artimage morphing, where an artimage is processed into a differently looking artimage. We obtain information from an object by measuring it and extracting knowledge from the information by analyzing it. However, we should also be aware that we are losing (or discarding) a huge amount of information in the mapping process. If we administer a math test, we obtain information (data) about math ability for the input "students"; however, this is practically synonymous to discarding (or neglecting) all information other than math ability. Similarly, if we analyze the math test data with IRT, we can obtain knowledge such as IRF and ability parameter; however, we lose all the other information.

As shown in the bottom part of Fig. 1.4, the more advanced the process, the higher the abstraction level of the information induced from the artimage, which is a merit of the artimage morphing. Every single object (phenomenon, event, or initial input) is normally too complex for direct comprehension. Therefore, we must first replace the input "students" with artimage "data" through testing for increasing levels of abstraction; however, this array of numbers is still extremely complicated and beyond direct interpretation. Therefore, we pass "data" through a statistical model to a further

---

[5] *Genzo* (現像) in Japanese, which is a compound word of "phenomenon" (現象) and "artificial" (イ).

increasing level of abstraction by changing it into an understandable artimage (i.e., knowledge).

Art

The concept of projection is widely applicable. For example, when an artist paints a landscape, or when a poet composes a poem about love, this can be expressed as follows:

$$f_{\text{PAINTER}}(\text{SCENERY}) = \text{PICTURE}, \quad \text{or} \quad \text{SCENERY} \xmapsto{f_{\text{PAINTER}}} \text{PICTURE},$$
$$f_{\text{POET}}(\text{LOVE}) = \text{POEM}, \quad \text{or} \quad \text{LOVE} \xmapsto{f_{\text{POET}}} \text{POEM}.$$

The picture or poem is considered to be an artimage processed by the artist or poet, respectively; however, in these cases, we observe that f is filled with the experience and emotion of the painter or poet; therefore, it is not only a process of information abstraction (or impoverishment) but also a process of information enrichment.

If the attributes of a student are enumerated successively, the number of attributes can easily exceed a million or billion; thus, each student is a high-dimensional object. However, of these approximately ten subjects can be measured by tests, such as languages, math, sciences, and social studies, and each test contains not more than 100 items. Why do we inquire into only 1000 items from the ten subjects of a student with a billion attributes? The test administrator should be accountable for why and how the 1000 items were selected, and a higher-stakes test[6] must be more responsible for this point.

Furthermore, even though there are many methods for analyzing data, we usually apply only a few. Why do we select those methods that produce indicators for each student and evaluate the student using the indicators that could determine the future of the student? The test administrator should also be accountable for this point.

Though we are given a variety of choices for measuring and analyzing (see Fig. 1.5), we can only trace a few routes in this tree; thus, we should be able to explain why we chose this process.

1.4 Test Data Engineer

As seen above, administering a test, obtaining data, and obtaining the results is a process of retouching an object into an artimage after another, which is generally expressed as

[6] A high-stakes test is a competitive test such that a high scorer on the test will lead an individual to a high social status (e.g., advanced civil service exam, medical licensing exam, bar exam, and entrance exams for leading universities).