2025

Missing Data in IRT & Psychometrics

Shenghai Dai

Washington State University

WASHINGTON STATE UNIVERSITY

Missing Data Terms

➢Types of Missing Data

- unit nonresponse - occurs when the entire data collection procedure fails on the participant
- Item nonresponse - partial data are available for the participant
- Wave nonresponse (also attrition, dropout) – Longitudinal data

➢Patterns of Missing Data

- univariate pattern
- monotone pattern
- arbitrary pattern

➢Mechanisms (or Distributions) of Missing Data

- **Missing completely at random (MCAR)**
- **Missing at random (MAR)**
- **Missing not at random (MNAR)**

Missing Data Patterns

univariate pattern

monotone pattern

arbitrary pattern

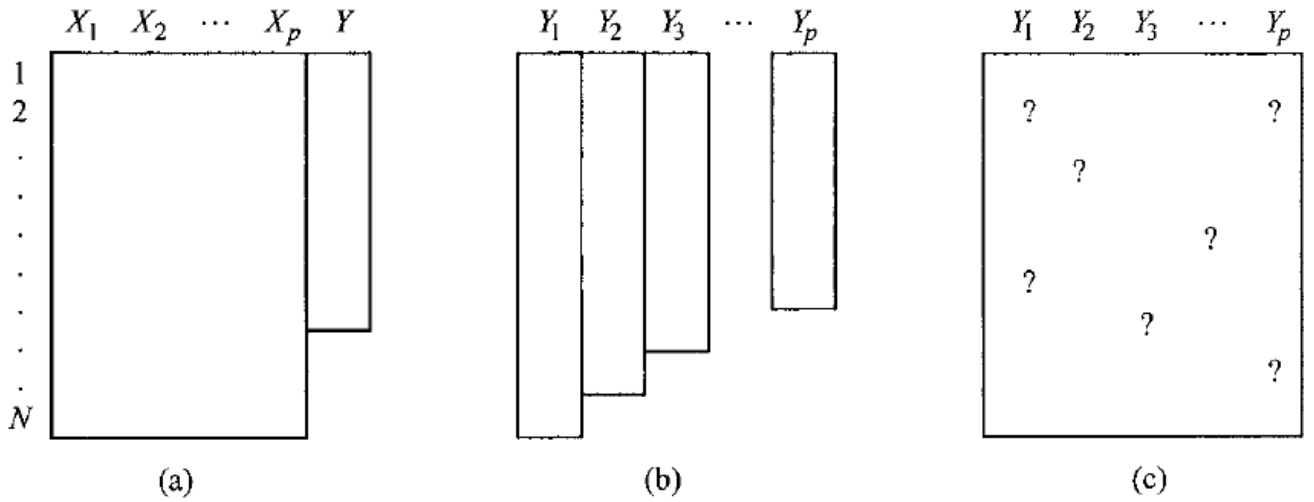


Figure 1. Patterns of nonresponse in rectangular data sets: (a) univariate pattern, (b) monotone pattern, and (c) arbitrary pattern. In each case, rows correspond to observational units and columns correspond to variables.

Schafer & Graham (2002)

Mechanisms of Missing Responses

- Little & Rubin (1987); Rubin (1976) etc.
- MCAR $P(R|Y_{com}) = P(R)$.
  - No system cause, random sample of complete data
  - Example: Missing-by-design
- MAR $P(R/Y_{com}) = P(R|Y_{obs})$.
  - MAR allows the probabilities of missingness to depend on observed data but not on missing data.
  - Caused by other measured variables but not the underlying level of the target trait – if observed, then MAR is also called ignorable missingness.
  - Example: missing caused by motivation on a math test
- MNAR
  - Depends on the missing value itself
  - Example: high-income people may be less likely to reveal their income
  - Nonignorable missingness

Mechanisms of Missing Responses

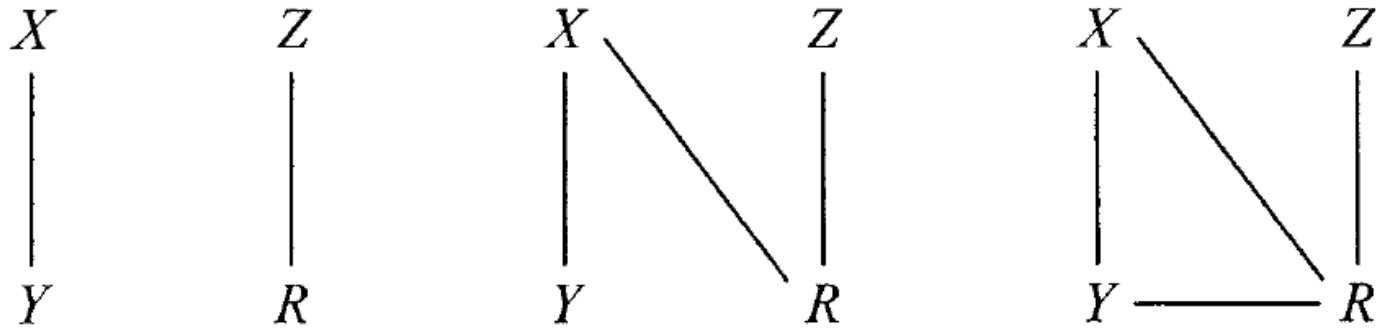


Figure 2. Graphical representations of (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) in a univariate missing-data pattern. X represents variables that are completely observed, Y represents a variable that is partly missing, Z represents the component of the causes of missingness unrelated to X and Y, and R represents the missingness.

Schafer & Graham (2002)

Types of Missing Responses

1. **Missing-by-design**
   - Result of booklet assessment designs
   - Usually treated as ignorable
2. **Omitted**
3. **Not reached**

Schafer & Graham (2002)

Omitted and Not-Reached Responses

1. **Omitted**
   - Missingness identified before the last valid response observed on an examinee's response sheet
   - Assumes that an examinee is presented with an item but skips it over either purposefully or accidentally.
2. **Not reached**
   - The consecutive missingness at the end of an examinee's response sheet, as it would occur if the examinee does not answer an item and any of the subsequent items in an assessment.

Coding For Omitted & Not-Reached

Note. Different assessment programs may operationalize the specifications slightly differently, mostly lying in **the coding of the missing response right after the last answer responded by the examinee.**

For example:

1. TIMSS: Omitted response
2. NAEP: not-reached unless
   - the item is the last one in a block and of the extended constructed-response (CR) format; and
   - the examinee responded to the preceding item

Missing Responses in Assessment Data

NAEP - Summaries of Item-Level Response Rates for NAEP Assessments
https://nces.ed.gov/nationsreportcard/tdw/analysis/initial_response.aspx

Grade	Statistic	Multiple-choice items	Dichotomous items	Polytomous items
4	Number of items	105	18	31
	Average percent missing[1]	2.65	6.83	4.21
	Minimum	.51	.99	.73
	Maximum	11.45	38.06	15.85
	Average percent off-task[2]	†	.08	.04
	Minimum	†	#	#
	Maximum	†	.27	.13
	Average weighted proportion correct	61.16	47.16	46.42

Missing Responses in Assessment Data

PISA 2018

Table 9.12

Average proportion correct, not-reached, and omitted for MSAT reading by stag

	Percent Correct		% Not-Reached		% Omit	
	Design A	Design B	Design A	Design B	Design A	Design B
Core Items	57.73	57.85	0.09	0.27	3.70	3.57
Stage 1 Items	55.15	47.51	1.84	12.57	5.49	7.42
Stage 2 Items	47.80	54.83	12.29	1.73	6.19	5.17

Note: Design A stage 1 and design B stage 2 items are the same items (highlighted); Design A stage 2 and design B stage 1 items are the same items.

Why Missing Responses Occur?

Table 2.1: Examples of reasons for missingness for combinations of intentional/unintentional missing data with item/unit nonresponse.

	Intentional	**Unintentional**
Unit nonresponse	Sampling	Refusal
		Self-selection
Item nonresponse	Matrix sampling	Skip question
	Branching	Coding error

Table 2.1 cross-classifies both distinctions, and provides some typical examples in each of the four cells. The distinction between intentional/unintentional missing data is the more important one. The item/unit nonresponse distinction says *how much* information is missing, while the distinction between intentional and unintentional missing data says *why* some information is missing. Knowing the reasons why data are incomplete is a first step toward the solution.

Source: van Buuren (2018) https://stefvanbuuren.name/fimd/sec-idconcepts.html

Why Missing Responses Occur?

Examinees characteristics?

- lack of self-confidence
- test-strategy
- knowledge of the answer
- student proficiency level
- item difficulty
- advertently skipping an item
- …

Why Missing Responses Occur?

Examinees characteristics?

- Student ability 3%
- Gender +0.4%
- ELL + 0.1%
- SES - 0.6%
- School type + 0%
- Race +0%
- School locale +0.1%
- IEP + 0.1%

Brown, Dai, Svetina (2014)

Why Missing Responses Occur?

Test items?

- Item difficulty 5%
- Position in block +0.7%
- Item complexity +1%
- Item format +13%
- Content subscale +3%

Brown, Dai, Svetina (2014)

Why Missing Responses Occur?

Mixed Findings

- **Pohl et al. (2014)**
  - Omitted Responses: examinees with a lower ability tended to omit more items
  - not-reached responses:
    - ✓ more not-reached responses from highly able examinees in **reading**
    - ✓ more not-reached responses from examinees with lower ability in **math**
- **de Ayala et al. (2001)**
  - examinees with higher ability might have a greater probability to leave an item unanswered when they did not know the answer than their peers with lower ability.
- **Robitzsch (2020) [and also some of Lord's work 1974, etc.]**
  - "This practically means that in the IRT model, observed item responses can be used to impute missing item responses. Since the probability Ppi(qp,M) is always larger than zero, ignoring missing item responses always leads to larger estimated trait values qp than scoring them as incorrect."

Coding Missingness

- **Omitted Responses**
  - Incorrect
  - Fractionally correct (e.g., ¼ if the item has four response options; oftenly in NAEP for MC item responses)
  - Missing
- **Not Reach Responses**
  - Incorrect
  - Missing/Not Administered

Note. **Coding varies across assessment programs and procedures. For example.**

TIMSS 2019

- **Omitted responses – always incorrect**
- **Not-reached responses – as missing in item calibration while as incorrect in estimating plausible values.**

Coding Missingness - NAEP

- In almost all NAEP Item Response Theory (IRT) analyses, missing responses at the end of a block of items are considered *not reached items* and are treated as if they had not been presented to the respondent. Occasionally, extended constructed-response items are the last item in a block. Because considerably more effort is required of the student to answer these items, nonresponse to an extended constructed-response item at the end of a block is considered an intentional omission (and scored as the lowest category) unless the student also did not respond to the item immediately preceding that item. In that case, the extended constructed-response item considered not reached is treated as if it had not been presented to the student. In the case of the national main and state writing assessment, there is a single constructed-response item in each separately-timed block. In the writing assessment when a student does not respond to the item or when the student provides an off-task response, the response also is treated as if the item had not been administered.
- Missing responses to items before the last observed response in a block are considered intentional omissions. If the **omitted item** is a multiple-choice item, the missing response is treated as fractionally correct at the value of the reciprocal of the number of response alternatives. If the omitted item is not a multiple-choice item, the missing response is scored so that the response is in the lowest category.
- These conventions are discussed by Mislevy and Wu (1988). With regard to the handling of not-reached items, Mislevy and Wu found that ignoring not-reached items introduces slight biases into item parameter estimation when not-reached items are present and speed is correlated with ability. With regard to omissions, they found that the method described above provides consistent limited-information maximum likelihood estimates of item and ability parameters under the assumption that respondents omit only if they can do no better than respond randomly.

Coding Missingness - PISA

Item response categories included several types of non-responses and item score categories. An item response was recoded as not-reached when a student did not answer the item or any subsequent item in the cluster for non-adaptive domains (mathematics, science, FL, and GC) or in the MSAT session for reading. An item response that did not perform properly in the field or had a missing human-coded response code was also converted to not-reached. An item response was recoded as omitted when a student did not answer the item but answered one or more of the subsequent items in the cluster or the MSAT reading form. The category off-task was used to identify an invalid missing category when a student did not answer the question in the expected way (e.g., by giving a response not associated with the item or responding with more than one answer in an exclusive choice question). In the computation of the item statistics and in the scaling analyses, the not-reached responses were excluded (i.e., treated as missing/ not-administered), but the omitted and off-task responses were treated as incorrect.

https://www.oecd.org/pisa/data/pisa2018technicalreport/Ch.09-Scaling-PISA-Data.pdf

Coding Missingness – **Context Scales**

TIMSS

- All cases with valid responses to at least **two** items on a scale were included in the calibration and scoring processes.

PISA

- For each scale, only persons with a minimum number of **three** valid responses were included.

IRT – Finally!

Cognitive Item Responses

- TIMSS 2019
  - ✓ 2- & 3-PL Models for dichotomous items
  - ✓ GPCM for polytomous items

$$P\left(x_i=1\middle|\theta_k,a_i,b_i,c_i\right)=c_i+\frac{1-c_i}{1+\exp\left(-1.7\cdot a_i\cdot\left(\theta_k-b_i\right)\right)}\equiv P_{i,1}\left(\theta_k\right)$$

$$P\left(x_i=l\middle|\theta_k,a_i,b_i,d_{i,1},\cdots,d_{i,m_i-1}\right)=\frac{\exp\left(\sum_{v=0}^{l}1.7\cdot a_i\cdot\left(\theta_k-b_i+d_{i,v}\right)\right)}{\sum_{g=0}^{m_i-1}\exp\left(\sum_{v=0}^{g}1.7\cdot a_i\cdot\left(\theta_k-b_i+d_{i,v}\right)\right)}=P_{i,l}\left(\theta_k\right)$$

- PISA (after 2015)
  - ✓ 2PL & GPCM with multi-group (i.e., country-by-language groups) IRT concurrent calibration

Context Questionnaire scales

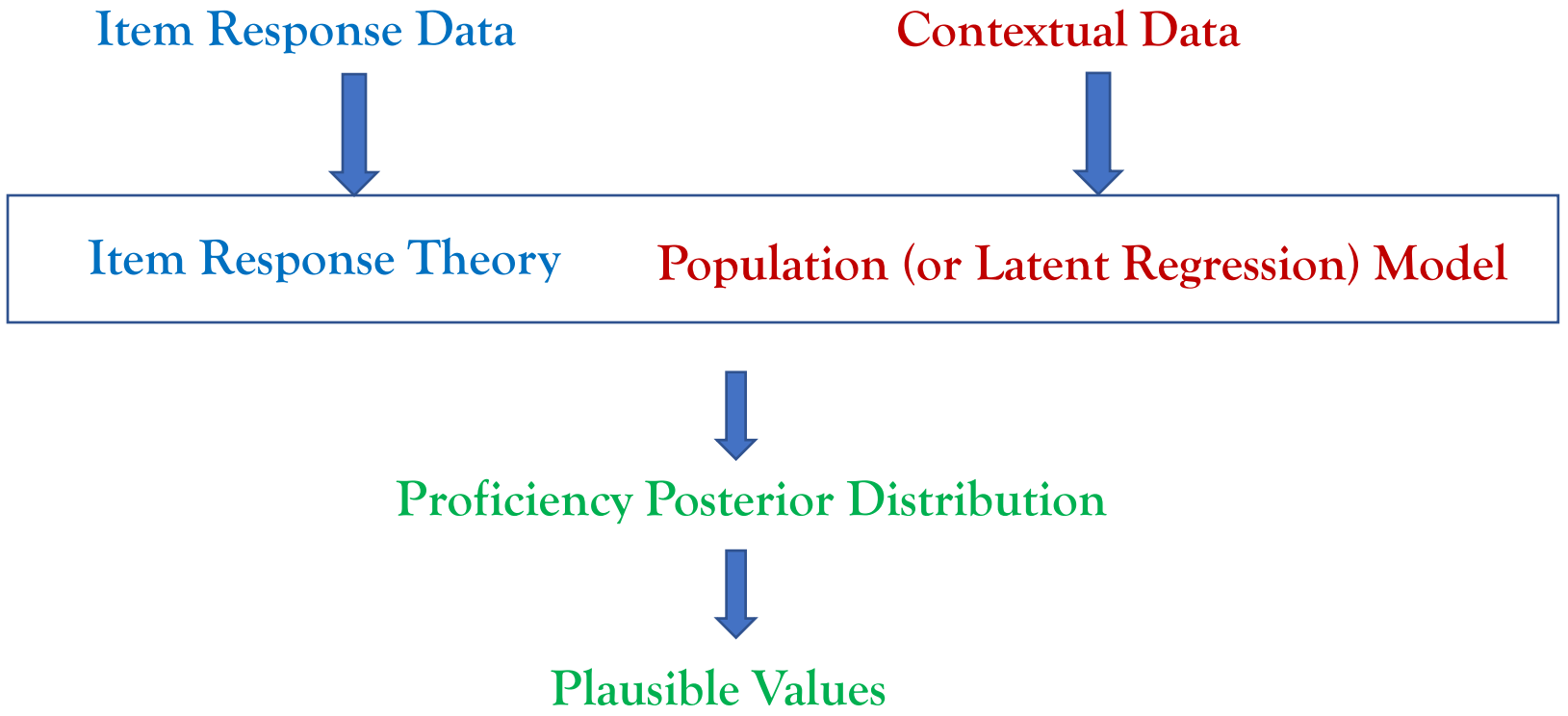
- TIMSS 2019 – PCM
- PISA 2022 (& after 2015) – 2PL & GPCM

$$P_i(x|\theta_n)=\frac{e^{\sum_{j=0}^{x}\left(\theta_n-\delta_i+\tau_{ij}\right)}}{\sum_{h=0}^{m_i}e^{\sum_{j=0}^{h}\left(\theta_n-\delta_i+\tau_{ij}\right)}}\qquad x=0,\,1,\,\ldots,m_i$$

$$P(x_{iv}=1|\theta_v,\beta_i,\alpha_i)=\frac{\exp\left(D\alpha_i(\theta_v-\beta_i)\right)}{1+\exp\left(D\alpha_i(\theta_v-\beta_i)\right)}$$

$$P\left(X_{ji}=k|\theta_j,\beta_i,\alpha_i,d_i\right)=\frac{exp\left(\sum_{r=0}^{k}D\alpha_i\left(\theta_j-\left(\beta_i+d_{ir}\right)\right)\right)}{\sum_{u=0}^{m_i}exp\left(\sum_{r=0}^{u}D\alpha_i\left(\theta_j-\left(\beta_i+d_{ir}\right)\right)\right)}$$

IRT Scaling in Large-Scale Assessments



Handling Missingness in IRT Modeling

- If missing data are **MCAR**, they can be treated as a random sample of the original complete observed data (i.e., no missing data is present) and ignored without introducing bias to parameter estimates.
  - Little's MCAR test (in SPSS)
  - Other tests such as *t*-test by using missing indicator as the grouping variable.
- What if missing data are **MAR?**
  - "MAR is a sufficient condition for pure likelihood and Bayesian inferences to be valid without modeling the missing mechanism." Little & Rubin (2001, p.14)
  - When a likelihood-function-based approach is used, MAR is treated as ignorable missingness, e.g., FIML in Mplus.
- What if missing data are **MNAR?**
- **Can we tell if missing data follows MCAR, MAR, or MNAR?**
  - **The answer is oftentimes NO!**

Handling Missingness in IRT Modeling

- Ignoring
- Single imputation
- Multiple imputation
- Model-base treatment

Ignoring Missing Responses

- **Listwise deletion** (LW, also known as complete case analysis)
  - Assumes MCAR
  - Oftentimes default for CTT and descriptive stats in software (SPSS, flexMIRT, IRTPRO, etc.) but not for IRT.
- **Pairwise deletion** (available-case analysis)
  - Assumes MCAR
  - Often paired up with limited information estimation methods such as the diagonally weighted least squares (DWLS) when the analysis is conducted under the ordinal factor analytic framework.
  - Default in Mplus for polychoric correlations when DWLS.
- Full information maximum likelihood (**FIML**) estimation
  - Assumes MAR
  - Default for IRT modeling in many software (flexMIRT, R ltm, etc.)

Single Imputation - I

- **Treating Missing Responses as Incorrect (IN)**
- **Treating Missing Data as Fractionally Correct (FR)**
- **Mean imputation or substitution** (Bernaards & Sijtsma, 2000)
  - Variable/Item mean (IM)
  - Person mean (PM)
- **Corrected mean (CM) imputation or substitution**
  - Corrected item mean (CIM; Huisman, 1999)
  - Two-way imputation (TW, Bernaards & Sijtsma, 2000)
  - TW-adj (Robitzsch & Rupp, 2009)
  - Bayesian-based TW with data augmentation (TW-DA, Ginkel et al., 2007)

Can be both **deterministic** vs. **stochastic:** PM-E, IM-E, CM-E, and TW-E

Single Imputation - II

➢Response Function (RF, Sijtsma and van der Ark, 2003)

- Assume item response function (IRF) for θ, but does not assume any item parameters
- Steps (for examine *i* and item *j*):

1. Let $\hat{R}_{(-j)i}$ be the rest score of examinee *i* on all available items except *j*, and *J* be the total number of items on a test. Then, $\hat{R}_{(-j)i} = \bar{y}_{i.}\,(J-1)$.
2. Define $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$ as the probability of endorsing a correct response for examinee *i* on item *j* based on the integer value of $\hat{R}_{(-j)i}$. Thus, if $\hat{R}_{(-j)i}$ is an integer, $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$ is the fraction of examinees with $\hat{R}_{(-j)i}$ who answer item *j* correctly. If $\hat{R}_{(-j)i}$ is not an integer, $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$ is computed by using its left and right neighbors.
3. Impute the missing response with a random draw from the Bernoulli distribution defined by $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$.

Single Imputation - III

- Expectation-Maximization (**EM**) imputation
  - Assumes MAR
- **Regression-based** imputation
  - Linear regression
  - Logistic regression
- **Other** methods
  - predictive mean matching (PMM)
  - machine learning methods

Multiple Imputation

- Developed by Donald B. Rubin.

Imputing one value for a missing datum cannot be correct in general, because we don't know what value to impute with certainty (if we did, it wouldn't be missing).

— Donald B. Rubin

- Now accepted as the best (?) general method to deal with missing data in many fields.

Steps of MI

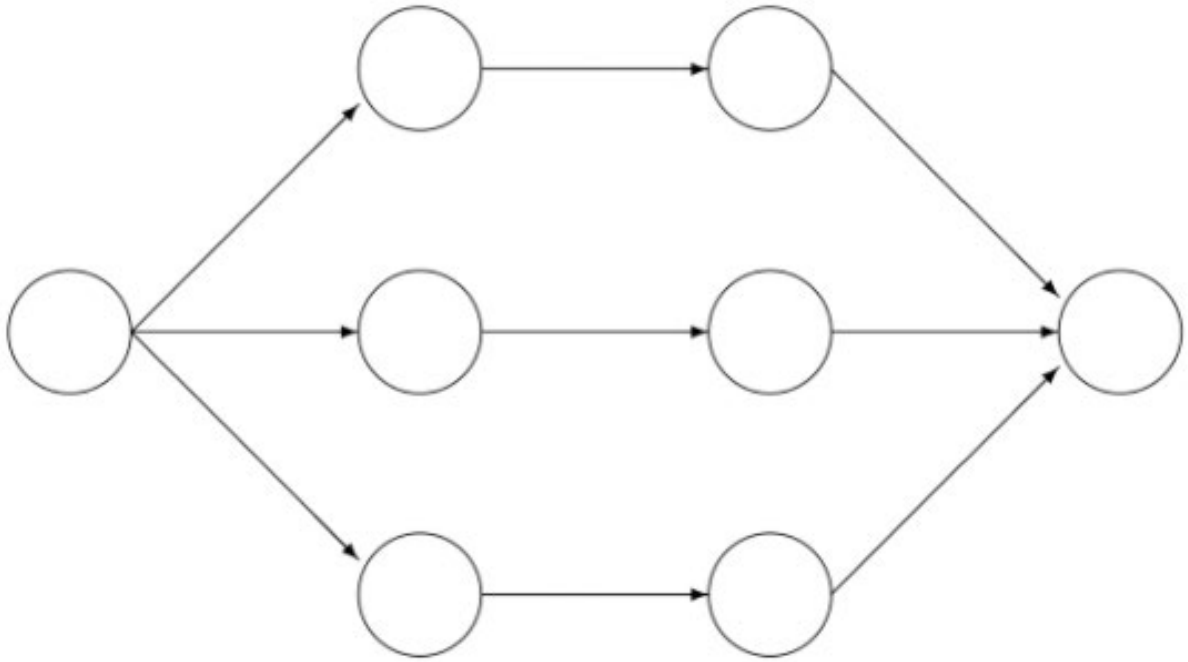


Figure 1.6: Scheme of main steps in multiple imputation.

Source: van Buuren (2018) https://stefvanbuuren.name/fimd/sec-nutshell.html

Steps of MI

➢**Step 1: Imputation**

- Select a specific imputation method: Any stochastic methods, MCMC, PMM, EM, etc.
- Impute the incomplete data m times. Usually $m = 5$.

➢**Step 2: Analysis**

- The same analysis conducted for each imputed data set.

➢**Step3: Pooling**

- A specific process (e.g., Rubin's rule) is used to pool the results by taking into account both the between and within-imputation variance

Approaches of MI

- **Joint modeling (JM)**
  - assumes that the data follow a joint probability distribution (e.g., the multivariate normal distribution) and the imputed values are drawn under the distribution.
- **Fully conditional specification (FCS)**
  - Also referred to as the multivariate imputation by chained equations **(MICE)** or sequential regression multiple imputation.
  - It does not assume a joint distribution but specifies a separate model for each variable and imputes for the missing values variable by variable.
  - different imputation models are specified based on the distribution and scale of the variables.
    - Ex. SPSS – logistic regression for categorical variables and linear regression for continuous variables.
  - MICE with classification and regression trees (MICE-CART) and the random forest imputation (MICE-RFI), the MI with latent class analysis (MILCA), etc.

Model-Based Methods - I

- Distributions of missing data are modeled in the analysis (through a missing indicator matrix). The underlying assumption of the methods is that examinees' missing tendencies are dependent on their ability.
  - **Missing tendency** (or response propensity, missing propensity, etc.)
    - ✓**Latent** – modeled simultaneously from a separate unidimensional IRT model; usually denoted as latent ignorability (e.g., Harel & Schafer, 2009)
    - ✓**Manifest** - computed as the relative number of missing responses across items (usually not recommended).
  - **Multiple-group IRT model or latent regression**
    - ✓Can handle both omitted and not-reached responses

Model-Based Methods - II

- **Missing tendency** (or response propensity, missing propensity, etc.)
- **Multiple-group IRT model or latent regression**
- **Challenges**
  - Unidimensionality underlying the missing indicators
  - Only capable of handling specific types of nonignorable missing mechanisms.
    - ✓Pohl and Becker (2020): the model-based approach "performs well when the missing mechanism is MCAR, MAR, or when the missing mechanism is generated according to the model for nonignorable missing values." (p. 2).
    - ✓That is, the approach may not be helpful if the non-ignorablility of the missingness does not rely on the missing tendency.
- **Readings (not assigned):** Holman & Glas (2005), Glas & Pimentel (2008), Rose et al. (2010), Pohl et al. (2014), Rose et al., 2017)

So… how good are these methods?

- Not a method stands out - mixed findings across studies
- A simulation study
  - Models: 2PL IRT, GRM
  - N = 1000
  - J = 20
  - Missing proportions: 15% vs. 30%
  - Missing data methods: LW, IN, PM, IM, TW, LR, EM, RF, and PMM

Dai (2021)

Results – 2PL IRT

Missing Rate	Missing Treatment	Item Discrimination		Item Difficulty	
		MAD	RMSE	MAD	RMSE
0	\	0.09	0.12	0.10	0.14
15%	LW	0.60	1.14	3.10	8.46
	IN	0.13	0.17	0.28	0.33
	PM	0.25	0.28	0.32	0.38
	IM	0.15	0.19	0.37	0.48
	TW	0.22	0.26	0.11	0.14
	LR	0.11	0.15	0.13	0.17
	EM	0.11	0.14	0.13	0.18
	RF	0.12	0.15	0.14	0.18
	PMM	0.11	0.15	0.13	0.17
	FIML	0.10	0.13	0.12	0.16
	mirt	0.11	0.15	0.13	0.17
30%	LW	\	\	\	\
	IN	0.16	0.21	0.58	0.71
	PM	0.55	0.58	0.54	0.63
	IM	0.25	0.31	0.90	1.20
	TW	0.48	0.53	0.13	0.16
	LR	0.14	0.18	0.16	0.22
	EM	0.15	0.19	0.18	0.25
	RF	0.16	0.21	0.25	0.31
	PMM	0.14	0.18	0.17	0.23
	FIML	0.12	0.15	0.14	0.19
	mirt	0.13	0.17	0.16	0.22

Results – GRM

Missing Rate	Missing Treatment	Item Discrimination		Item Thresholds							
				b1		b2		b3		b4	
		MAD	RMSE	MAD	RMSE	MAD	RMSE	MAD	RMSE	MAD	RMSE
0	\	0.06	0.08	0.10	0.13	0.07	0.09	0.07	0.09	0.10	0.13
15%	LW	0.30	0.38	1.41	3.67	0.72	1.64	0.63	1.56	1.37	4.23
	IN	0.14	0.17	0.50	0.54	0.33	0.35	0.29	0.32	0.33	0.38
	PM	0.16	0.18	0.14	0.17	0.10	0.12	0.07	0.09	0.09	0.12
	IM	0.19	0.21	0.64	0.69	0.55	0.58	0.33	0.36	0.44	0.49
	TW	0.14	0.16	0.16	0.19	0.09	0.12	0.07	0.09	0.09	0.12
	LR	0.08	0.10	0.16	0.21	0.10	0.13	0.08	0.10	0.12	0.15
	EM	0.09	0.11	0.24	0.30	0.13	0.16	0.07	0.10	0.14	0.19
	RF	0.13	0.15	0.26	0.32	0.16	0.20	0.16	0.20	0.25	0.31
	PMM	0.08	0.11	0.17	0.22	0.11	0.14	0.08	0.10	0.12	0.16
	FIML	0.07	0.09	0.14	0.19	0.09	0.12	0.08	0.10	0.11	0.14
	mirt	0.08	0.10	0.15	0.20	0.10	0.13	0.08	0.10	0.12	0.15
30%	LW	\	\	\	\	\	\	\	\	\	\
	IN	0.22	0.26	1.08	1.13	0.78	0.81	0.68	0.72	0.72	0.77
	PM	0.26	0.29	0.28	0.33	0.20	0.23	0.10	0.13	0.11	0.14
	IM	0.28	0.31	1.28	1.36	1.11	1.17	0.79	0.83	0.93	1.00
	TW	0.25	0.27	0.34	0.38	0.23	0.26	0.11	0.14	0.14	0.17
	LR	0.13	0.16	0.34	0.42	0.20	0.24	0.11	0.14	0.16	0.21
	EM	0.14	0.17	0.50	0.57	0.26	0.30	0.09	0.12	0.24	0.30
	RF	0.33	0.36	1.03	1.12	0.84	0.89	0.83	0.89	1.00	1.09
	PMM	0.14	0.17	0.36	0.44	0.22	0.26	0.12	0.15	0.16	0.22
	FIML	0.11	0.14	0.31	0.37	0.19	0.23	0.11	0.14	0.13	0.18
	mirt	0.12	0.15	0.33	0.41	0.20	0.25	0.11	0.14	0.15	0.20

Findings

- No method is optimal. To date, rules and guidelines remain unclear on the selection of an appropriate imputation method for missing responses.
- Different methods embody different assumptions about the mechanisms and distributions of the missing responses.
- The decision on the selection of a specific method should be made with caution and based on reasonable explanations.
- News methods have been proposed. Again, mixed findings.
- Some good methods: MI, model-based methods, FIML, PMM

Dai (2021)

Others – IRT & Missingness

➢ **Some findings/recommendations**

- The item parameter estimation accuracy was lower in the presence of a short instrument (J = 3) and/or a small sample size (N = 150), especially when items were of poor quality and the missing rate was 20% or higher.
- Generally, the impact of missing data was acceptable when MR = 10% or less. When the missing rate was high (MR ≥ 20%), a larger sample size of at least 300 and an instrument length of at least five items were required for acceptable item parameter estimations.
- The performance of GPCM was more stable than GRM across conditions, especially those of missing data. A large impact of missing data, especially when MR ≥ 30%, was detected on GRM in estimating person parameters.
- a rule of thumb from existing literature suggests a sample of 200–300 with an instrument length of 10 or more items for accurate parameter estimation. Similarly, results support that a sample size of at least 300 in the implementation of both GRM and GPCM.
- An N = 150 might be feasible when the purpose is to obtain the person parameters as it will lead to inaccurate item parameter estimations and inflated type II error rates for the model fit indices.

Dai et al., 2021

Others – IRT & Missingness

"It is important that one's approach to the treatment of omitted and not reached items take into account the specifics of the testing conditions and the audience for whom reports are intended. On one hand, some members of the educational measurement community are at ease discussing state-of-the-art developments in complex IRT modeling. On the other hand, other members in the same extended community are not comfortable with assessment results that can differ according to how missing data are treated—let alone the differences that occur depending on the particular IRT model employed. Common sense and reasonable explanations still must govern decisions regarding how test data are analyzed and reported."

Ludlow & O'leary (1999, p.629)

IRT & Missingness in Speeded Tests

- 1PL, 2PL, GPCM
- Missingness = 25%, 50%
- J = 10, 40
- N = 500, 1000
- Correlation between ability and missing propensity = 0, 0.2, 0.4, 0.6, 0.8
- "**Intermediate missing responses were regarded as ignorable. However, this may not always be realistic**. In these cases one may use the approach of Holman and Glas (2005) and apply the Rasch model as a model for the missing data indicator for intermediate missing responses. One can then either assume that these responses depend on a third latent proficiency variable, say, y3, that is correlated with the two other variables, or one may assume that these intermediate missing responses depend on the same latent proficiency variable that also describes the not-reached items through the steps model."

Glas & Pimentel (2008)

IRT & Missingness - More

- **Glas et al. (2015)** - polytomous response and response propensity models with covariates
- **Xiao & Bulut (2020)** - **FIML, IN,** MICE-CART, & MICE-RFI in 3PL
- **Zhang & Walker (2008)** – missingness on person fit and estimation with dichotomous items (2PL IRT)
  - "Consistent with previous findings (Glasser, 1964; Little, 1992), this study demonstrated that the pairwise deletion method is the optimal way to deal with missing data when assessing person–model fit. Moreover, the best way to recover person trait level is to treat missing data as omitted."
- **Finch (2008)** – compare 7 methods on IRT models
  - no method was superior to others in terms of item parameter recovery rates.
  - MI, listwise, and fractionally correct were preferred if the purpose was to achieve lower biased item parameter estimates

IRT & Missingness – Bayesian Related

- **Patz & Junker (1999) –** MCMC in IRT with missing data (and others)
- **van Ginkel et al. (2007) –** two-way imputation – a Bayesian method
- **Fu et al. (2014) –** Gibbs sampling in multidimensional logistic response model
- **Aβmann et al. (2015) –** Bayesian in IRT with missing in background variables

Missingness in MIRT

- **Bernaards & Sijtsma (1999)** compared 7 missing data methods on multidimensional data, and extended to 14 methods on factor analysis in **Bernaards & Sijtsma (2000).**
- **Andreis & Ferrari (2012)** – Forward imputation with NLPCA, miss forest, and MICE to handle missingness in MIRT
- **Bacci & Bartolucci (2014)** – multidimensional latent class IRT

Missingness in DIF Detection

- **Rousseau et al. (2006) -** compared MI, IN, as Missing in handling omitted items for MH, LR, NCDIF.
  - MH is more robust
- **Garrett (2009) –** Mantel and ordinal logistic regression were compared using within-person mean substitution and multiple imputation when data were missing completely at random. [PCM used to generate data].
  - Results indicated that the performance of the Mantel and ordinal logistic regression depended on the percent of missing data in the data set, the magnitude of DIF, and the sample size ratio.
- **Robitzsch & Rupp (2009) –** MH & LR; LW, TW, RF, IN;
  - "An incorrect treatment of missing data can thus lead to severe increases of Type I and Type II error rates. However, the choice between the two DIF detection methods investigated in this study is not important."
- **More:** Finch (2011), Goodman et al. (2011)

Missingness in CDA/CDM/DCM

- **Zhang (2013) -** missingness and skill mastery profiles of CDA
- **de Chiusole et al. (2015) –** missing data in knowledge space theory
- **Xu & von Davier (2006)** – GDM with missingness on NAEP
- **Dai et al. (2017) –** missingness on Q-matrix validation
- **Dai & Svetina Valdivia (2022) –** missingness on DINA

Useful (Free) Books

- Van Buuran (2018) **Flexible Imputation of Missing Data**
  - https://stefvanbuuren.name/fimd/

- Heymans and Eekhout (2019) **Applied Missing Data analysis using SPSS and R**
  - https://bookdown.org/mwheymans/bookmi/

Suggested Readings

- Dai, S. (2021). Handling missing responses in psychometrics: Methods and software. *Psych*, 3, 673-693. https://doi.org/10.3390/psych3040043
- Dai, S., Vo, T., Kehinde, O.J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of polytomous IRT models with rating scale data: An investigation over sample size, instrument length, and missing data. *Frontiers in Education – Assessment, Testing and Applied Measurement. 6, 721963*. https://www.frontiersin.org/articles/10.3389/feduc.2021.721963
- Enders, C. K. (2023). Missing data: An update on the state of the art. *Psychological Methods*.
- Ludlow, L. H., & O'leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement*, 59(4), 615–630.
- Mislevy, R. J., & Wu, P. K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*, *1996*(2), i–36. https://doi.org/10.1002/j.2333-8504.1996.tb01708.x
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4), 525–556.
- Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, 74(3), 423–452.
- Robitzsch, A. (2020). *About Still Nonignorable Consequences of (Partially) Ignoring Missing Item Responses in Large-scale Assessment*.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7(2), 147.

Thank you!

For questions, contact s.dai@wsu.edu

More information see https://labs.wsu.edu/lsd/