Alternative Reliability Measures for Medical Students' Skills in Objective Structured Clinical Examinations

Chi Chang, PhD, MS
Heather Laird-Fick, MD, MPH
Carol Park, PhD, MPH
David Solomon, PhD

College of Human Medicine
Michigan State University

chisq@msu.edu

SHARED DISCOVERY CURRICULUM

MICHIGAN STATE UNIVERSITY | College of Human Medicine

Outline and Disclaimer

- **Outline**
  - Study Background
  - Cognitive Diagnostic Model
    - Two reliabilities
  - Illustration dataset
  - Results
  - Discussion

- **Funding Source and Disclaimer**

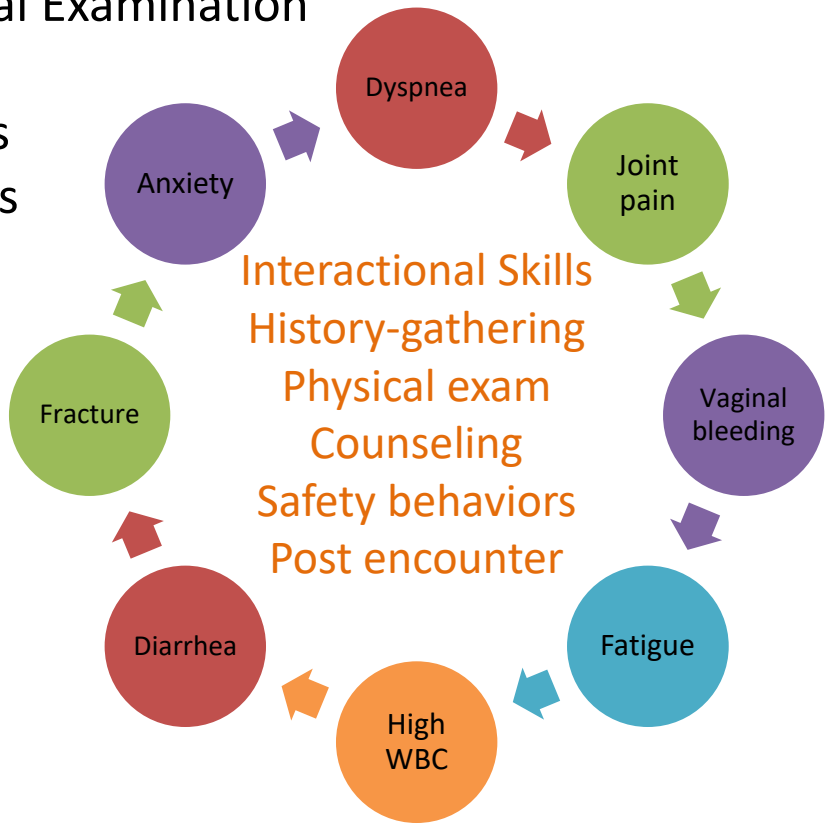
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Study Background

- **Study Purpose**
  - Provide two alternative reliability measures for skills in objective structured Clinical Examinations (OSCE)
    - Classification Accuracy
    - Classification Consistency
  - Illustrate these measures with an empirical sample

Study background

- Progress Clinical Skills Examination
  - 8-station objective structured clinical Examination
  - Standard patient checklists
  - Faculty-graded post encounter tasks
  - Twice per semester for 10 semesters
  - Criterion-referenced standards for performance



- Data structure
  - Cases
  - Domains
  - Skills
  - Items

- Grain size: Case > Domain > Skill > Item

Study Background

- Reliability
  - Definition: the degree to which test scores are precise and free from measurement error.
  - Generalizability Theory (G-study/G-theory)
    - Variance decomposition method
    - Students' score variations are attributable to multiple sources
      - e.g. case specificity, domains, checklist items, Standard patients.
    - Accounting for these multiple sources of variation would complicate the model.
    - Minimizing the complexity by aggregating scores across a certain course would compromise the information and decrease the capability of reliability indices.
    - Increasing the number of cases or checklist items would allow all the variance sources to be incorporated, this is expensive and impractical.

Introduction

- **Cognitive Diagnostic Model** (CDM)[1, 2]
  - **CDMs have been applied in**
    - mathematics education to diagnose students' difficulties in learning fraction questions[3, 4]
    - social anxiety disorder study to identify the subgroup of individuals with social phobias.[5]
  - Test Blueprint – **Q-matrix** (Item × skill matrix)
    - Item is denoted by $j, j = 1 \ldots J$
    - Skill is denoted by $k, k = 1 \ldots K$

	Add	Sub	Mul
$7 + 5 \times 2$	1	0	1
$6 - 3$	0	1	0

$$\text{item} \begin{bmatrix} q_{11} & \cdots & q_{1k} \\ \vdots & \ddots & \vdots \\ q_{J1} & \cdots & q_{JK} \end{bmatrix} \quad \text{skill}$$

- ✓ Each element of the Q-matrix is dichotomous.
- ✓ If item j requires skill k to be answered correctly, then q_{jk} = 1; otherwise, $q_{jk} = 0$

- Example: If there are only three skills we want to assess, the number of possible skill patterns is 8 (i.e., $2^3 = 8$), denoted by $\boldsymbol{\alpha}$,
  - [0, 0, 0]
  - [0, 0, 1]
  - [0, 1, 0]
  - [1, 0, 0]
  - [1, 1, 0]
  - [1, 0, 1]
  - [0, 1, 1]
  - [1, 1, 1]

Introduction

- **Model Specification**
  - the deterministic inputs, noisy "and" gate (DINA) Model
  - The probability for answering an item correctly requires an examinee who has all the necessary skills not to slip and an examinee who lacks at least one of the required skills to guess correctly.

$$P(X_{ij} = 1 | \alpha_{ik}, q_{jk}) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})}$$

  - X_{ij} denotes the observed score from examinee i, $i = 1 \ldots I$
  - $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$,
    - If the student possess all the required skills for the item: $\eta_{ij} = 1$; otherwise, $\eta_{ij} = 0$
  - Skill pattern is denoted by $\boldsymbol{\alpha}$: $\alpha_i = [\alpha_{i1}, \ldots, \alpha_{iK}]$
  - Skill pattern required for each item: $q_j = [q_{j1}, \ldots, q_{jK}]$

Introduction

- The purpose of CDM is to **identify examinees' skill pattern**.
- The estimated skills are **dichotomous** (possess or not).
- Reliability is the degree to which test scores are precise and free from measurement error.
- **Reliability:**
  - **Classification Accuracy:**
    - The proportion of students whose estimated classification memberships were matched with their **true skill classification membership**.
    - The probability that the estimated classification membership is equal to the true classification membership.
  - **Classification Consistency:** [6, 7]
    - The proportion of students whose estimated classifications were identical **across the two simulated parallel assessments**.
    - The probability that two parallel forms of the assessment result in the same estimated classification
      - Parallel forms of a CDA as two tests with the same Q-matrix and identical item parameters.

Methods

Data Source

- 190 second-year medical students' PCSE scores from the summer semester 2019
- 8 clinical scenario patient-encounter cases
- 15 minutes per case
- ~20 items per case were assessment by standard patients covering five domains:

Domain	# items	# skills
Interaction	32	3
History-gathering	42	5
Physical Examination	36	5
Counseling	31	5
Safety Behaviors	18	4

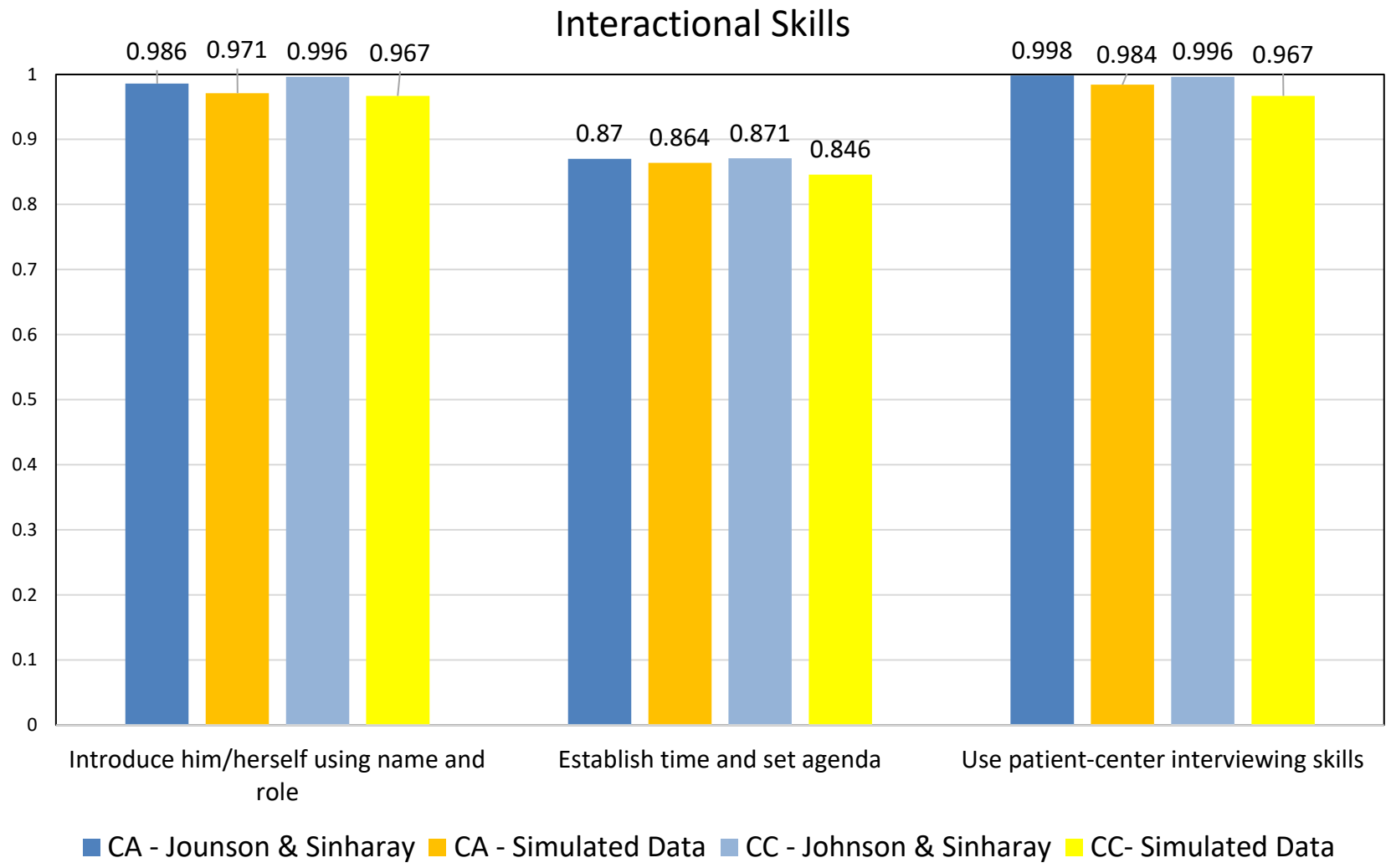
Table 2. Domains and associated skills used in Q-matrix

Domains	Skills
Interactional Skills (32 items total)	1) Introduce himself/herself using name and role (medical students) 2) Establish time and set agenda 3) Use patient-center interviewing skills
History-gathering domain (42 items total)	1) Elicit potential causes of symptoms or associated symptoms 2) Elicit health/disease management history 3) Seek important psychosocial information 4) Elicit basic description of symptoms 5) Obtain medical history
Physical Examination (36 items total)	1) Examine head, eyes, ears, nose, throat (HEENT) 2) Perform cardiopulmonary examination 3) Perform abdominal examination 4) Examine integument 5) Perform neuromusculoskeletal examination
Counseling Skills (31 items total)	1) Assess health beliefs 2) Provide patient-centered education 3) Indicate next steps in evaluation and management 4) Communicate possible causes of symptoms 5) Engage in shared decision making
Safety Behaviors (18 items total)	1) Verify patient using two identifiers 2) Identify risk 3) Perform hand hygiene 4) Elicit medication history

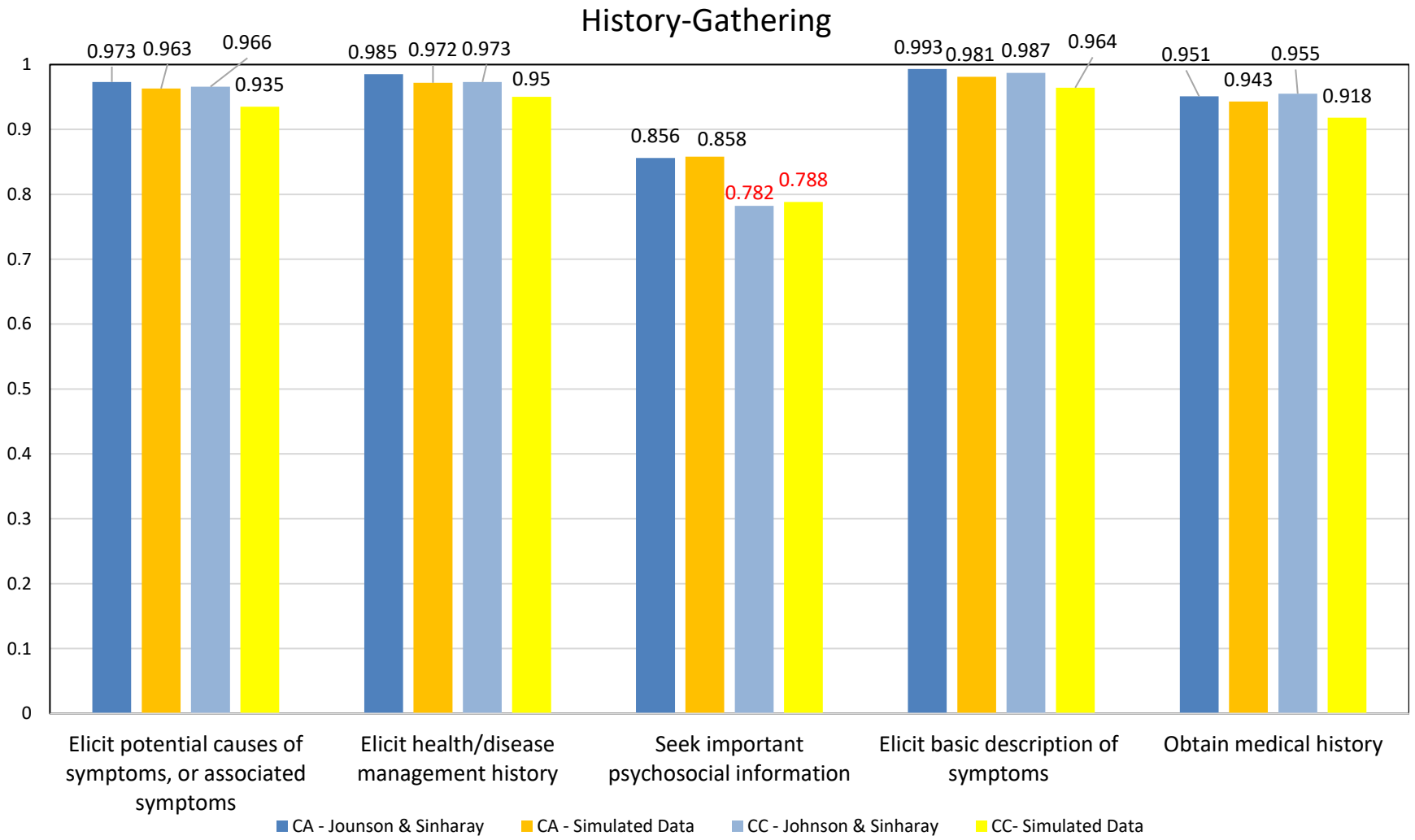
Methods

- **Model specification:** DINA model
- **Q-Matrix**: Clinical educators
- **Model identification**:
  - Each skill in the domain must be measured by at least 3 items
  - Each domain must have 5 skills or under.
- **Reliability Indices:**
  - Classification Accuracy & Classification Consistency
  - Both were estimated using Johnson & Sinharay[8] and simulated-based estimations.
  - Maximum a posteriori (MAP) estimators were used to decide the final membership estimates in Monte Carlo simulations.
    - the sample size is set to be 5,000 for each simulated dataset.
  - Four reliability indices were calculated at the skill-level and pattern level.
  - Cutoffs: .95 is excellent reliability, .90 very good, .80 good, and .7 fair.
- **Statistical software**:
  - R version 3.5.3, *CDM* package[9]

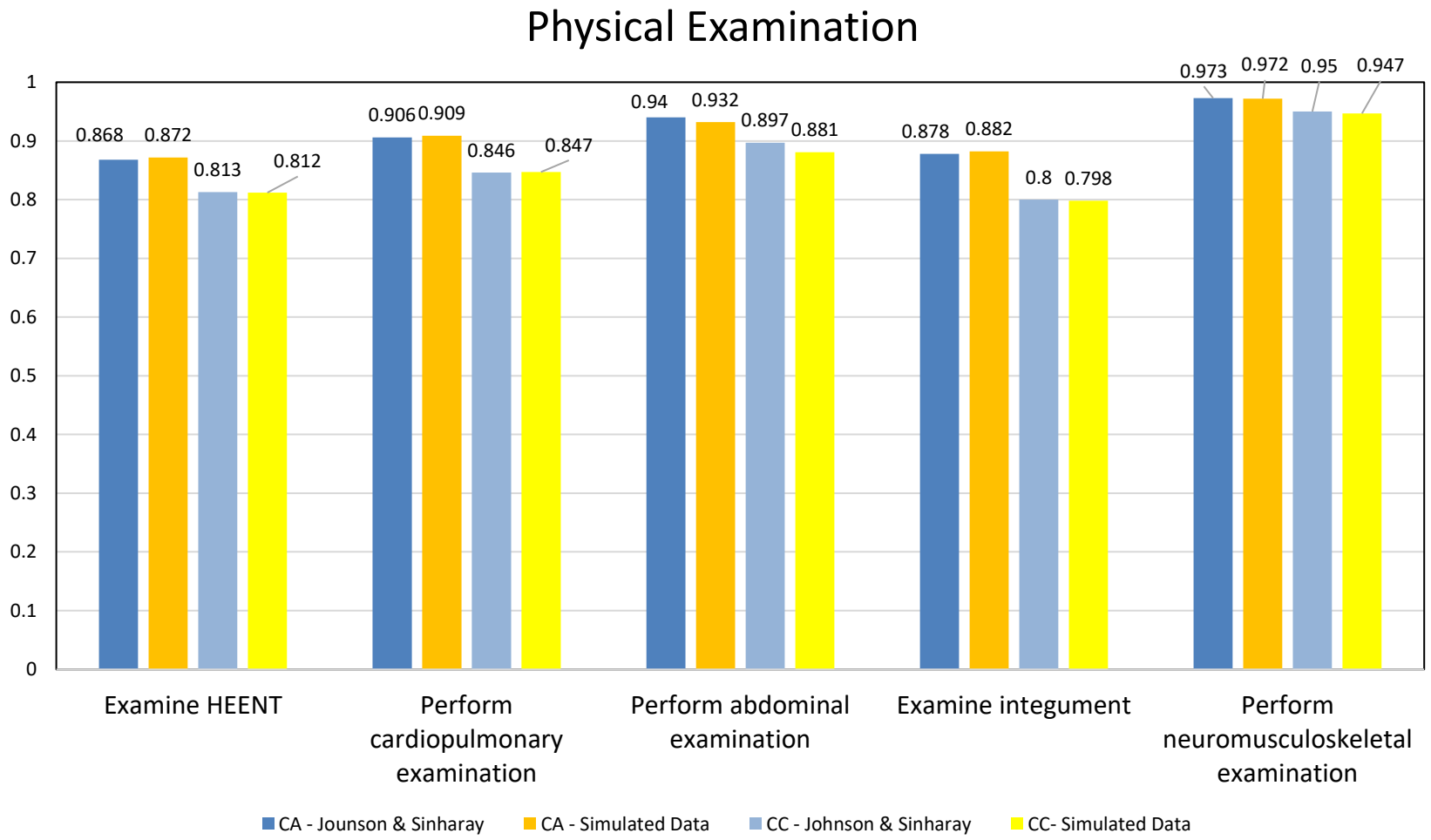
Results – Skill-level Reliability in the Interaction Domain



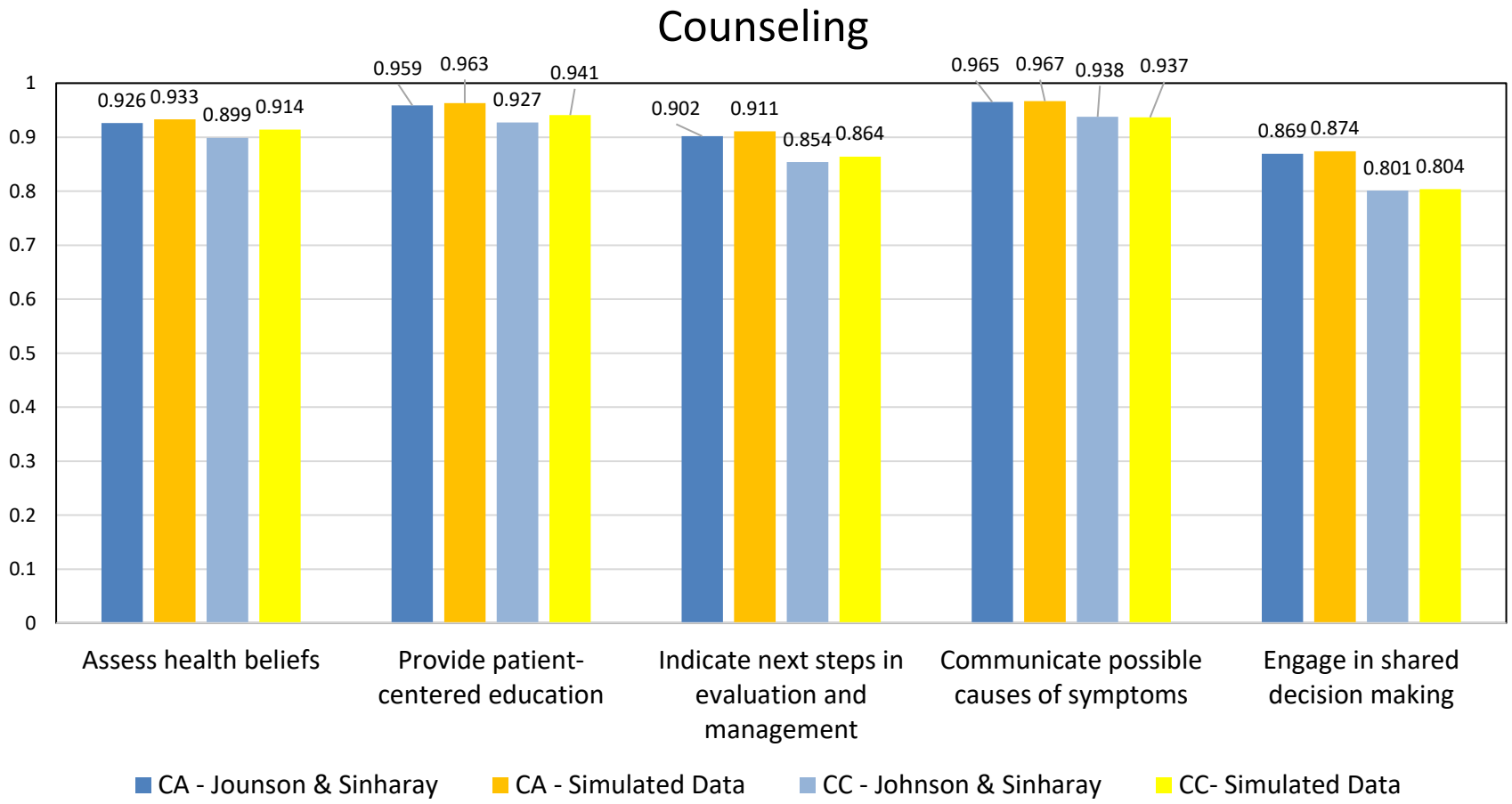
Results – Skill-level Reliability in the History-Gathering



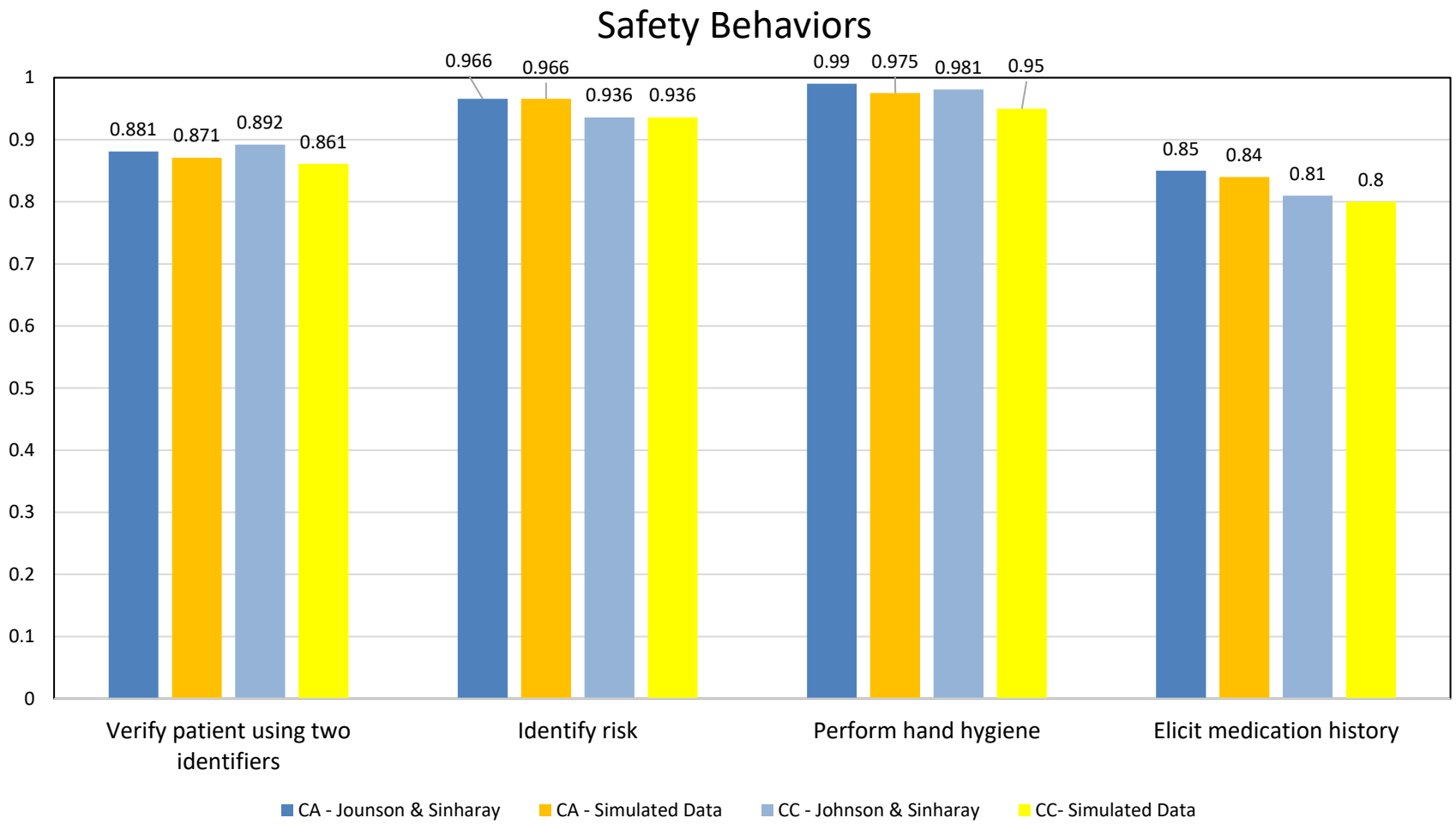
Results – Skill Reliability in the Physical Examination Domain



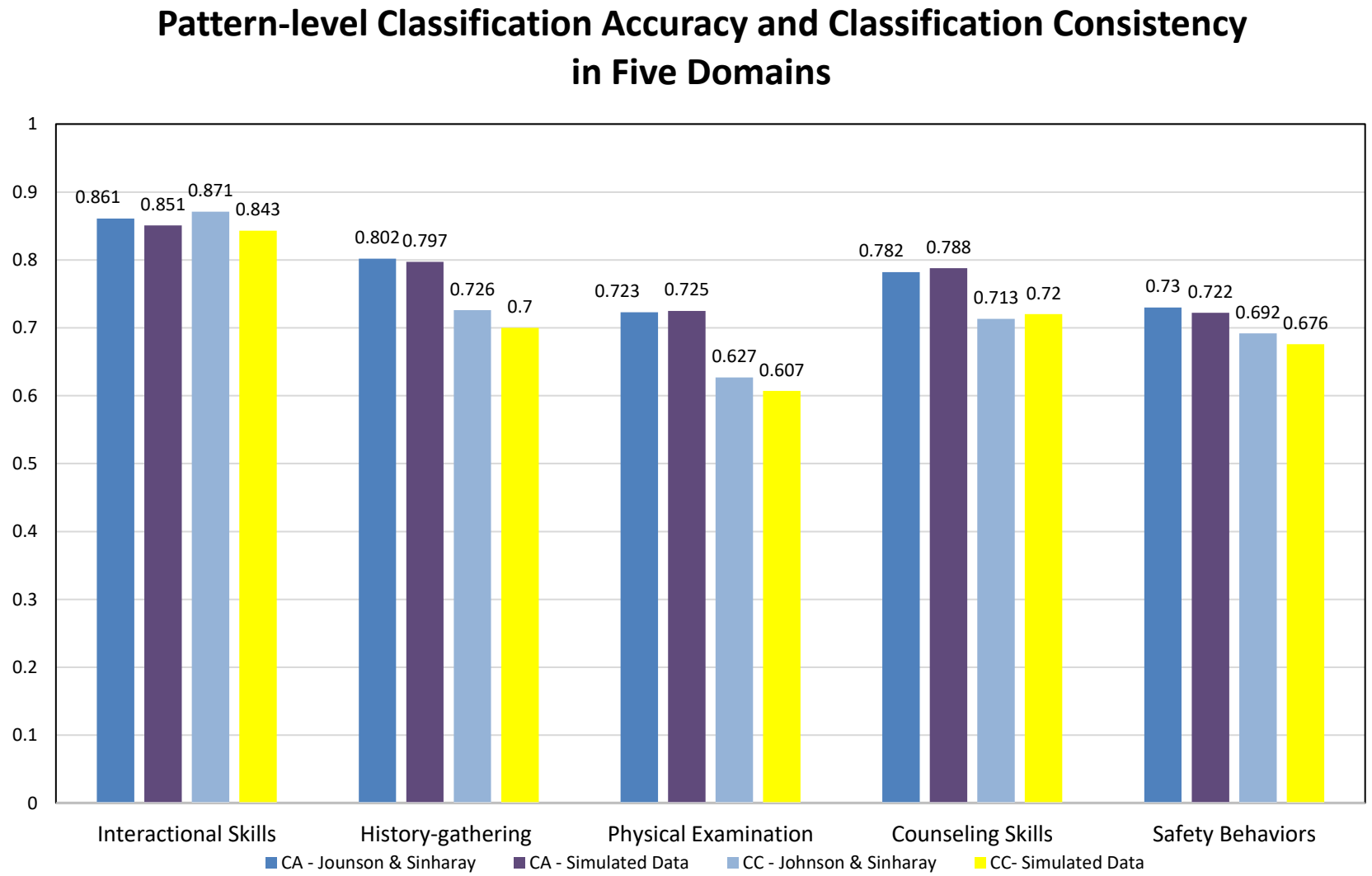
Results – Skill-level Reliability in the Counseling Domain



Results – Skill Reliability in the Safety Behaviors Domain



Results – Skill Reliability in the Safety Behaviors Domain



Conclusion

- **Conclusion**
  - Except for the skill "**Seek important psychosocial information**" in the History-gathering Domain, all skills had at least good reliability indices ($c > .8$).
  - Pattern-level reliability ranged from 0.607 to 0.871.

- **Next Step**
  - For a skill with low reliability, we can further explore whether if
    - SPs that need additional training in coding specific medical students' behavior
    - the description of the item that is too hard to understand
    - memory burden
  - For a domain with low pattern-level reliability, we can further explore which pattern is the most problematic.

Discussion

- **Applicability and Practicality**
  - Model students' learning level using the fine-grained information
  - Students and teachers can focus on specific unmastered skills to improve: Great to support classroom learning
  - Allow for small sample size and short test lengths
  - Available for skill-level reliability and pattern-level reliability
  - Bridge the gap between psychometricians and clinical educators.

- **Limitation and Future Study**
  - Model specifications
  - Q-matrix validation
  - Classification Validation
  - Polytomous item responses
  - Hierarchical skills

References

1. Leighton JP, Gierl MJ, Hunka SM. The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach. 41(3):205-237. doi:10.1111/j.1745-3984.2004.tb01163.x

2. Rupp AA, Templin J, Henson RA. *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: The Guilford Press; 2010.

3. Templin J, Jacobson E, Bradshaw L, Izs A. Diagnosing Teachers ' Understandings of Rational Numbers : Classification Framework. *Educ Meas Issues Pract*. 2014;33(1):2-14.

4. Tatsuoka C, Ferguson T. Sequential classification on partially ordered sets. *J R Stat Soc Ser B*. 2003;65(1):143-157. doi:10.1111/1467-9868.00377

5. Fang G, Liu J, Ying Z. On the identifiability of diagnostic classification models. *ArXiv*. 2017.

6. Cui Y, Gierl MJ, Chang H-H. Estimating classification consistency and accuracy for cognitive diagnostic assessment. *J Educ Meas*. 2012;49(1):19-38. doi:10.1111/j.1745-3984.2011.00158.x

7. Wang W, Song L, Chen P, Meng Y, Ding S. Attribute-Level and Pattern-Level Classification Consistency and Accuracy Indices for Cognitive Diagnostic Assessment. 52(4):457-476. doi:10.1111/jedm.12096

8. Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement, 45*(4), 635-664. 10.1111/jedm.12196

9. George AC, Robitzsch A, Kiefer T, Gross J, Ünlü A. The R Package *CDM* for Cognitive Diagnosis Models. 74(1):1-24.

Thank you!