

MSCEIT meets MIRT

Assessing the Psychometric Properties of Emotional Intelligence

Matthew Sigal

February 23rd, 2015

The MSCEIT

The **Myers-Solovey-Caruso Emotional Intelligence Test** (MSCEIT)

- Has been in publication in various forms since 1990
- Is one of the most famous measures of Emotional Intelligence
- Primary authors:
 - Dr. J. Mayer (Professor of Psychology, University of New Hampshire)
 - Dr. P. Salovey (President of Yale University)
 - Dr. D. Caruso (Co-Founder of the EI Skills Group)

The MSCEIT

- An “ability” based measure of Emotional Intelligence (EI)
- 141¹ 5-level Likert-type items, which provides:
 - One **Total Score**
 - Two **Area Scores** (Experiential, Strategic)
 - Four **Branch Scores** (Perceiving, Using, Understanding, Managing)
 - Eight **Task Scores**
 - Three **Supplemental Scores** (Scatter, Positive/Negative Bias)
- In the literature:
 - Overall scores and branch scores have been shown by some researchers to correlate with a variety of outcomes.
 - Other researchers have failed to find expected relationships.

¹Of the 141 items, only 122 are actually scored.

Revising the MSCEIT

- Last revision to the test was published in 2002
- The authors, in collaboration with Multi-Health Systems, have decided it was time for another revision
- Changes will primarily be dictated by issues raised by previous research and analysis of MSCEIT customer database

Here we go...

I was given:

- an administration of the test
- a brief introduction what it is supposed to measure, alongside the current manual, and item bank
- a *large* dataset ($N = 104,496$), with participants from North America, Asia, Latin America, Oceania, and Europe.

Here we go...

I was given:

- an administration of the test
- a brief introduction what it is supposed to measure, alongside the current manual, and item bank
- a *large* dataset ($N = 104,496$), with participants from North America, Asia, Latin America, Oceania, and Europe.

...and asked: *"Please confirm the factor structure of the MSCEIT"*.

Guiding Principal

Ensure appropriate results across the entire instrument,
not just at the overall and scale levels...

This entailed being critical of:

- **External considerations** (such as regional/ethnic biases)
- The **structure** of the test
 - How the MSCEIT is **scored**
 - The MSCEIT's **item content**

...and provide feedback to the authors and the Research and Development Team in charge of the revision regarding my findings.

Sample Proportions

	Asian	Black	Hispanic	White	Total
North America	7,509	4,702	4,086	49,581	65,878
Asia	3,943	24	38	11,85	5,190
Latin America	167	92	2,693	178	3,130
Oceania	1,459	35	145	15,846	17,485
Europe	1,104	198	220	11,291	12,813
Total	14,182	5,051	7,182	78,081	104,496

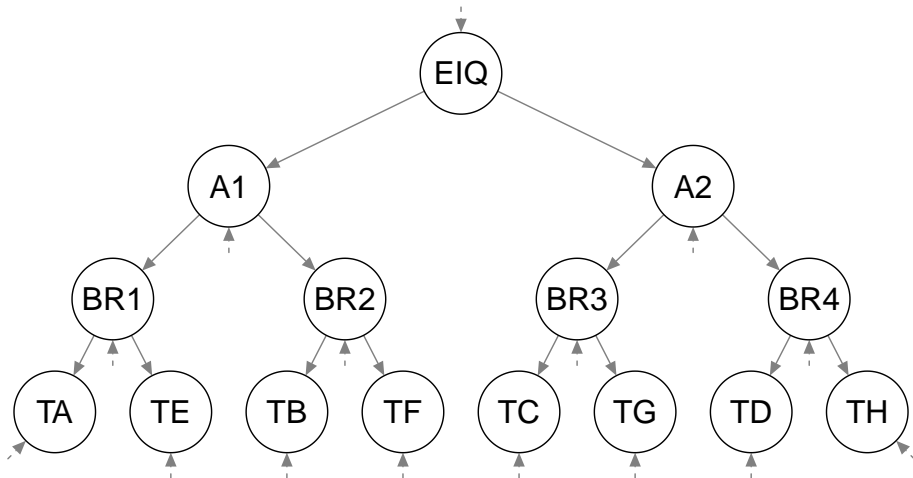
Factor Analysis

What does the MSCEIT look like?

The primary research goal from R&D pertained to the factor structure of the MSCEIT. In their literature review, they found some researchers did not like the 4 branch model, and showed a preference instead for a 3 branch model. Others found that area scores were unnecessary.

Overall, given this large sample size, could we ascertain the "correct" factor structure model for the MSCEIT?

Factor Analysis: Structural Model



Scoring

The model is estimable only with the imposition of drastic equality constraints to make it identifiable. However, this was further complicated by an examination of the scoring methods used for the MSCEIT:

- Scores for each item are given based upon **consensus scoring**, or how well a respondent's key corresponds to a panel of experts or against a "General Population" normative sample.
 - The General Population sample was predominately white
 - ... and scores are weighed to match the 2000 US census (70% white)
- Task and Scale scores are then calculated using item level parcels²
 - Which are then converted to empirical percentiles
 - And then standardized to have a mean of 100 and *SD* of 15

²For 6 of the 8 tasks, parcels were created due to common item stems, but for 2 of the tasks they were created ad hoc.

Scoring

A goal of the revision is to reduce possible ethnic and regional biases, not replicate them.

- Put aside the original scoring algorithms
- Return to the **raw data**
- Base subsequent analyses on the categorical response patterns at the item level

Regional Differences

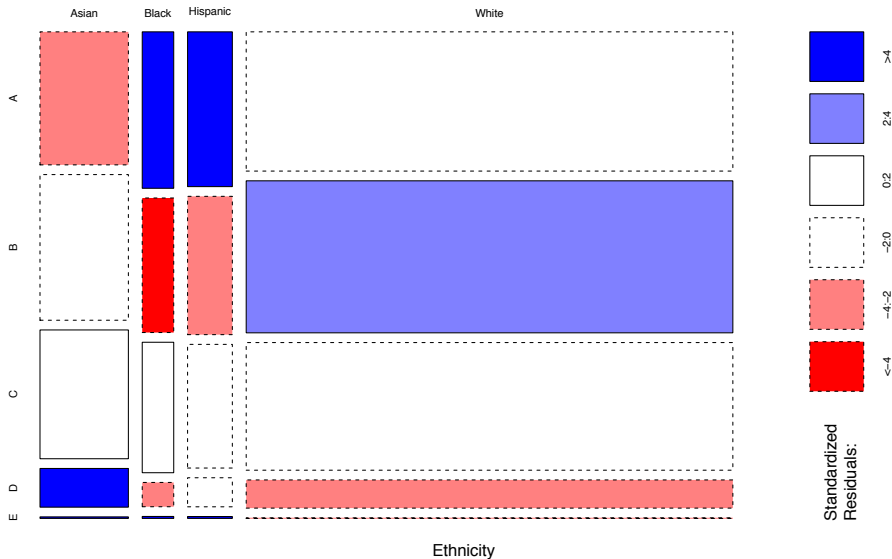
- **Numerically:** Interested in contingency tables
 - But we have a very large sample size
 - So much data, small deviations are significant
 - Could concentrate on effect sizes instead (ϕ over χ^2)
- **Visually:** Look at patterns of response via mosaic plots

Regional Differences

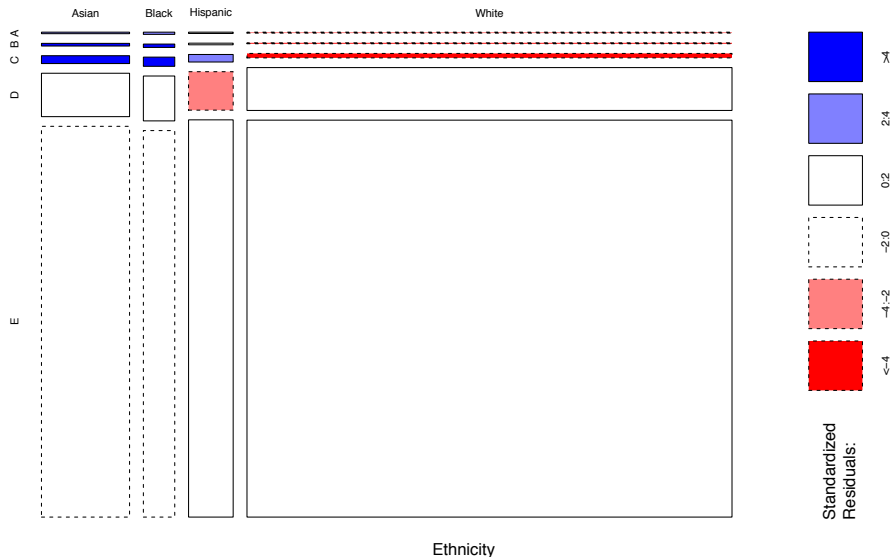
On the raw data, this could pertain to:

- Full contingency table looking at all three variables simultaneously:
 - $5 \text{ (responses)} \times 4 \text{ (ethnicities)} \times 5 \text{ (regions)} = 100 \text{ cells for each item!}$
 - This is too complex to visually inspect
- Item response by ethnicity *across* region
 - e.g., do the ethnic groups respond differently within North America?
- item response by region *within* ethnicity
 - e.g., do Asians from North America respond different than Asians in Latin America?

Item 2: Response Options by Ethnicity



Item 23: Response Options by Ethnicity



Regional Differences

Unfortunately. . .

- Comparing hundreds of mosaic plots isn't very illuminating; and,
- Comparing ϕ values is not a perfect solution, as it is interpreted in relation to a variable maximum ($k - 1$, where k is `min(nrow, ncol)`)
 - This value changes if not all response options were used, or if we are looking at particular combinations of region/race
- Also, neither approach shows how respondents were answering in comparison to the expert or the general population normative samples.

Flattened Data: Coherent Item

RESPONSE		A	B	C	D	E
Task A	Expert	100.00	0.00	0.00	0.00	0.00
Item 11	GenPop	87.72	9.69	1.86	0.49	0.25
North America	Asian	74.40	20.87	3.52	0.95	0.27
	Black	77.07	16.80	4.74	1.11	0.28
	Hispanic	79.66	16.18	3.13	0.69	0.34
	White	76.90	19.42	3.15	0.42	0.12
Asia	Asian	68.35	22.57	6.54	1.70	0.84
	Black	83.33	12.50	0.00	4.17	0.00
	Hispanic	73.68	26.32	0.00	0.00	0.00
	White	72.66	21.69	4.90	0.59	0.17
Latin America	Asian	86.83	10.18	2.40	0.00	0.60
	Black	84.78	11.96	3.26	0.00	0.00
	Hispanic	84.48	12.37	2.56	0.19	0.41
	White	70.23	25.84	3.93	0.00	0.00
Oceania	Asian	68.68	25.98	4.18	0.82	0.34
	Black	74.29	20.00	5.71	0.00	0.00
	Hispanic	67.59	26.21	4.14	2.07	0.00
	White	71.03	24.65	3.77	0.48	0.07
Europe	Asian	69.11	23.10	5.98	1.18	0.63
	Black	67.17	24.24	6.06	2.02	0.51
	Hispanic	77.73	18.18	3.18	0.91	0.00
	White	72.99	22.61	3.56	0.61	0.23

Flattened Data: Discoherent Item

RESPONSE		A	B	C	D	E
Task D Item 4	Expert	52.38	42.86	0.00	0.00	4.76
	GenPop	4.45	15.93	18.33	32.44	28.87
North America	Asian	5.22	20.48	19.42	30.27	24.61
	Black	4.68	15.02	18.38	29.99	31.94
	Hispanic	5.04	14.05	17.43	31.91	31.57
	White	4.23	22.41	21.29	32.93	19.14
Asia	Asian	4.36	13.85	15.65	35.94	30.21
	Black	4.17	25.00	29.17	37.50	4.17
	Hispanic	5.26	7.90	15.79	36.84	34.21
	White	6.92	30.80	17.55	27.00	17.72
Latin America	Asian	4.19	10.78	20.36	29.34	35.33
	Black	4.35	15.22	29.35	28.26	22.83
	Hispanic	2.41	11.59	12.55	27.70	45.75
	White	5.62	18.54	16.29	34.27	25.28
Oceania	Asian	4.59	21.11	19.53	30.23	24.54
	Black	0.00	17.14	11.43	40.00	31.43
	Hispanic	4.83	22.07	17.93	27.59	27.59
	White	4.44	28.03	20.59	32.69	14.25
Europe	Asian	4.17	19.11	20.47	32.88	23.37
	Black	1.52	18.69	17.68	30.30	31.82
	Hispanic	4.09	21.36	12.27	25.46	36.82
	White	5.16	24.86	19.75	32.03	18.21

Follow-up Analyses

- With the assistance of R&D, each item was evaluated in terms of its response proportions across region, ethnicity, and scoring method.
- Items patterns that seemed unreasonable were identified and flagged.
- Of the 19 items that did not contribute to the original scoring:
 - 17 were flagged again here.
- Of the 122 scored items:
 - 17 items were flagged.

The remaining items served as the item pool
for further exploratory analysis.

Subsequent Analysis

As the data I am interested in is categorical and the questions are of variable usefulness, I wanted to model it using item response theory (IRT).

IRT

IRT is a set of statistical methods used for assessing item level contributions to a latent variable model. Most often used in context of traditional (dichotomous) scoring methods, and usually applied on unidimensional tests.

General idea: a person's **ability** underlies each of their test responses, and will affect the proportion to which various responses are endorsed.

Item Response Theory

Unlike with consensus scoring, we want to see which items only **high ability** individuals get correct (high difficulty) versus the items which **everyone gets correct** (low difficulty).

IRT can generally be thought of as the intersection between factor analysis and logistic regression.

Item Response Theory

Unlike the data used in traditional IRT models, the current analysis posed a few additional problems:

- ① There is no obvious “correct” answer for each question
- ② 6 of the 8 scales featured “testlets”, or items with common stems
- ③ The large amount of data makes running models computationally intensive

The Nominal IRT Response Model

To deal with the first issue, I turned to the **Nominal Response Model** (NRM; Bock 1972, 1997), as implemented in R in the `mirt` package (Chalmers, 2012).

- The NRM frees us from relying on particular constraints imposed by other scoring rubrics
 - e.g., one alternative, the Graded Response Model (GRM), assumes that there is a correct response and each category further from that response is less and less correct in a strictly ordinal fashion
- Allows us to test the assumptions/feasibility of the GRM
- Exploratory procedure for use in cases where a meaningful scoring key is not available

The Nominal IRT Response Model

GOAL

Utilize item response patterns, in conjunction with empirically derived “high” and “low” categories, to assess item contributions to test information.

The Nominal IRT Response Model

Given some ability, θ , the probability of endorsing a particular response option (k), is non-linearly related to item easiness (d), and it's slope/discrimination (a).

$$P(x = k|\theta) = \frac{\exp(ak_k*(a\theta)+d_k)}{\sum_{j=1}^k \exp(ak_k*(a\theta)+d_k)}$$

This model has **3 sets of parameters**:

- 1 The **ak_k parameters** are scoring/ordering coefficients for the k th response option, and are used to investigate the empirical ordering of the response options
 - For identification purposes, one response category is constrained to 0 and another to $k - 1$
 - The larger the value, the better it is at discriminating higher level abilities

The Nominal IRT Response Model

$$P(x = k|\theta) = \frac{\exp(ak_k*(a\theta)+d_k)}{\sum_{j=1}^k \exp(ak_j*(a\theta)+d_j)}$$

- ② The **a parameter** indicates how discriminating the question is on the underlying factor;
- ③ The **d_k parameters** are the relative estimated response proportions for each response compared to the first response category, d_0 , which is constrained to 0.

The Bifactor Nominal Response Model

The **traditional NRM** is unidimensional – all items load on one factor.

The **bifactor NRM** is an extension that allows us to account for interdependencies within items, by allowing for testlet factors while also allowing all items to load on a common underlying dimension.

If we have 4 testlets, we will have 5 a parameters, with a_1 indicating the item loading on the underlying factor, and a_2 - a_5 indicating the loadings for each item on their testlet.

Application to the MSCEIT

Six of the eight tasks on the MSCEIT feature testlet-style items and were assessed using the bifactor NRM.

Items on the remaining two tasks are independent, and were evaluated with the NRM.

Analysis Details:

- As the authors believe each task targets a particular aspect of EI, analyses were conducted on a task by task basis.
- To alleviate concerns over missing data, only complete cases were used.
- Initial models concentrated on just the North American subset.
- The NRM is an *exploratory* model, and the parameters are sensitive to where the ak constraints are placed. If any ak value is greater than $\text{abs}(10)$, it indicates a poor choice for the fixed ak values, and ak for the largest value should be changed to $(k-1)$.

Example: Task A (Faces), Item 2 (Fear)

a1	a2	a3	a4	a5
0.59	0.39	0.00	0.00	0.00

Large loading on primary factor

Positive moderate loading on testlet factor

ak0	ak1	ak2	ak3	ak4
4.00	0.84	-0.78	-0.35	0.00

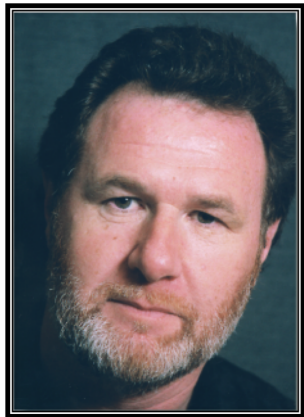
Responses ranged from 1 (No) to 5 (Extreme)

Response 1 has largest scoring coefficient

$ak0 > ak1 > ak2$, but $ak3$ and $ak4$ differ.

d0	d1	d2	d3	d4
0.00	0.92	0.30	-1.00	-4.36

Response 2 is estimated to be endorsed most frequently, followed by 3, then 1.




Again, what are we looking for?

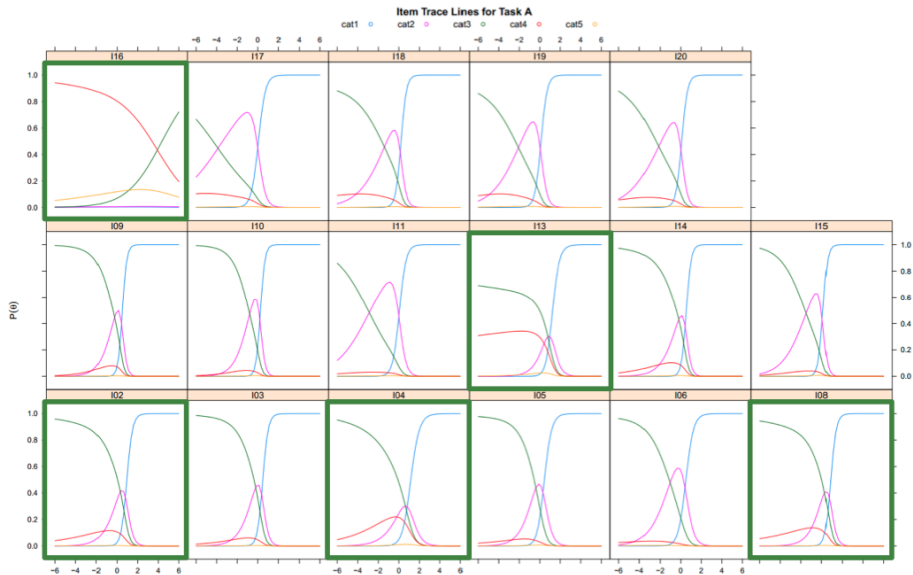
Primary concerns:

- Magnitude of a parameters (strength of factor loadings)
- Ordering of ak parameters (if truly ordinal, $ak_0 < ak_1 < \dots < ak_4$)
- Magnitude of d parameters (dispersal of responses across distractors)

Secondary concerns:

- Model fit statistics (AIC, BIC, Log-Likelihood)
- Empirical reliability statistics
- SS Loadings

		a1	a2	a3	a4	a5	ak0	ak1	ak2	ak3	ak4	d0	d1	d2	d3	d4
	Fear	0.59	0.39	0.00	0.00	0.00	4.00	0.84	-0.78	-0.35	0.00	0.00	0.92	0.30	-1.00	-4.36
	Surprise	0.57	0.31	0.00	0.00	0.00	4.00	0.67	-1.88	-0.92	0.00	0.00	-0.78	-2.68	-4.03	-6.11
	Disgust	0.30	0.26	0.00	0.00	0.00	4.00	0.39	-2.04	-1.28	0.00	0.00	0.68	0.57	0.07	-2.50
	Excitement	0.34	0.07	0.00	0.00	0.00	4.00	0.50	-2.64	-1.57	0.00	0.00	-0.75	-2.35	-4.24	-5.11
	Happiness	0.45	0.00	0.28	0.00	0.00	4.00	1.06	-0.95	-0.87	0.00	0.00	-0.51	-2.59	-5.19	-6.85
	Fear	0.57	0.00	-0.11	0.00	0.00	4.00	0.54	-1.16	-0.99	0.00	0.00	0.57	0.00	-1.31	-4.11
	Surprise	0.63	0.00	0.14	0.00	0.00	4.00	0.21	-1.73	-0.84	0.00	0.00	-0.84	-2.52	-3.47	-5.68
	Excitement	1.37	0.00	1.03	0.00	0.00	4.00	-0.71	-3.01	-1.84	0.00	0.00	-4.92	-11.22	-10.15	-8.36
	Happiness	0.35	0.00	0.00	0.18	0.00	4.00	-0.19	-2.36	-1.07	0.00	0.00	-1.93	-4.70	-5.84	-6.67
	Fear	0.48	0.00	0.00	-0.01	0.00	4.00	1.29	-0.31	-0.54	0.00	0.00	0.94	1.18	0.63	-1.53
	Surprise	0.61	0.00	0.00	0.25	0.00	4.00	0.41	-1.10	-0.75	0.00	0.00	-0.74	-2.00	-3.21	-5.26
	Excitement	1.11	0.00	0.00	0.93	0.00	4.00	-0.79	-2.67	-1.50	0.00	0.00	-5.52	-10.73	-9.84	-8.50
	Happiness	0.24	0.00	0.00	0.00	-0.32	3.56	4.00	3.68	0.94	0.00	0.00	0.47	3.51	5.91	4.01
	Sadness	0.45	0.00	0.00	0.00	0.58	4.00	0.76	-1.27	-1.43	0.00	0.00	-2.23	-6.13	-7.15	-7.54
	Fear	0.66	0.00	0.00	0.00	0.67	4.00	0.80	-0.63	-1.26	0.00	0.00	-2.32	-5.01	-7.59	-7.74
	Anger	0.93	0.00	0.00	0.00	1.29	4.00	1.40	-0.15	-0.76	0.00	0.00	-3.25	-8.15	-11.41	-10.85
	Disgust	0.79	0.00	0.00	0.00	1.00	4.00	1.07	-0.54	-1.11	0.00	0.00	-3.28	-7.53	-10.37	-9.33



Statistical Output

Statistic	Bifactor Model	Unidimensional Model
Log-likelihood	-921,577	-949,100
AIC	1,843,460	1,898,473
AICc	1,843,461	1,898,473
BIC	1,844,852	1,899,710
SABIC	1,844,366	1,899,278

Statistic	F1	F2	F3	F4	F5	F1
Empirical Reliability	.85	.33	.37	.33	.57	.88
SS Loadings	1.8	.10	.21	.20	.76	1.2

Outcomes of Analysis

- Compared information statistics between unidimensional NRM and bifactor NRM
 - Bifactor fit substantially better than unidimensional for 5/6 tasks
- Looked at SS Loadings to get idea of which items “define” the model
- Found that the “correct” response for many items was on one extreme or the other
 - On many questions, middle option was chosen with large proportion
 - Possible cause: Seeing 3 as “neutral”, rather than average amount.
- 2 tasks have responses that are independent
 - The NRM empirically selected the response that the experts chose as “most correct” the vast majority of the time.

Outcomes of Analysis

Via discussions with authors and R&D, the revision will:

- Address the identifiability of the structural model:
 - Incorporate new tasks to provide a minimum of three tasks per branch
 - Test for models without “Area” level scores, and examine performance of the three branch model
- Address item level concerns:
 - Rewrite substantial portion of item bank to focus more on identifying presence of emotion, rather than its absence
 - Remove items that were shown to exhibit bias across regional groups
 - Use partial credit method for scoring, based upon a diverse expert panel
 - Rework item response labels to help ensure clarity

Future Work

- Utilize IRT results to trim unhelpful items from test
- Utilize Differential Test Functioning to assess structural patterns across Ethnic and Regional groups
- Help R&D with the evaluation of new items and tasks