テスト理論の概要と実践

加藤 健太郎
ベネッセ教育総合研究所

2024.9.9

教育協力ウィーク2024 『JICA算数KMNプレゼンツ テストの開発の第一人者に迫る：テスト理論の概要と活用～JICA算数学習到達度評価（J-MAGPA）の実践を通して～』 発表資料

本発表の内容・構成

内容

テスト性能の評価の規範となるテスト理論・妥当性理論の概要をJICA算数学習到達度評価（J-MAGPA）の分析例を示しながらお話しします

構成

1. テストによる測定の基本的なアイデア
2. 古典的テスト理論モデルとテストの信頼性

   古典的テスト理論にもとづくJ-MAGPAの分析結果
3. 妥当性とその検証

   J-MAGPAの妥当性検証の観点

テストによる測定の基本的なアイデア

テストとは？

「・・・能力，学力，性格，行動などの個人や集団の特性を測定するための用具であり・・・心理学的なテスト，学力・知識試験はもとより，行動評定，態度評定などの評定手法，調査のほか，構造化された面接，組織的観察記録にも・・・」（『テスト・スタンダード』，p. 17）

日本テスト学会（編） (2007). テスト・スタンダード―日本のテストの将来に向けて 金子書房

AERA/APA/NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

（暗黙の前提？）テストは測定ツールの１つに過ぎない。多数の人の特性を，客観的に，公平に，かつ効率的に測定する際に威力を発揮する。

「測る」ということ

一定のルールに従って，対象や事象の持つ何らかの属性に対して数値を割り当てること

武蔵工業大学（編） (2004). なんでも測定団が行く 講談社ブルーバックス
Joseph, C. (Ed.) (2005). *A measure of everything*. Firefly Books.

物理量の測定

長さ，距離，質量，温度，色，成分，時間，カウント etc.

測定対象が実在する（実在概念と見なせる）

対象を直接（物理的に）測定することができる

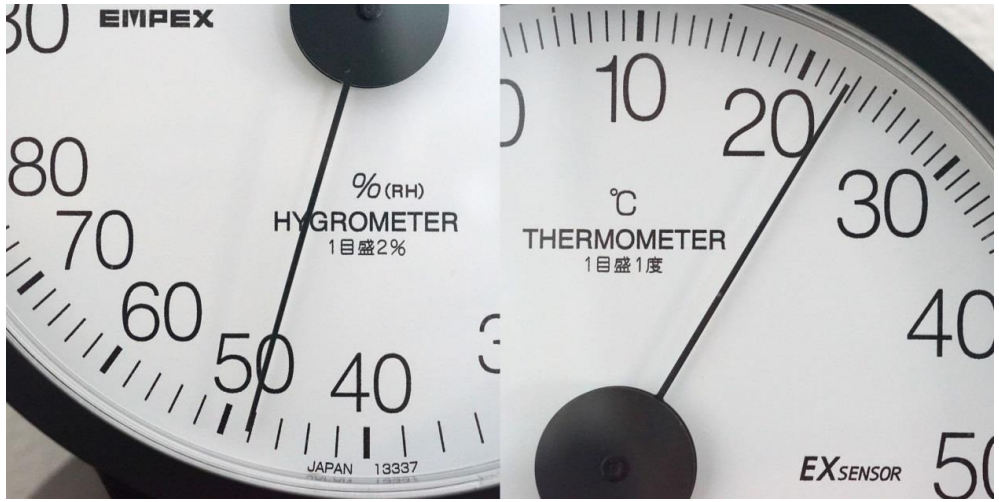
　　尺度（原点・単位）や測定値の解釈が明解

対象を繰り返し測定することができる

測定対象は１つ（その都度，特定の対象がある）

測定には誤差が伴う

温湿度計の例



製品仕様
温度計：バイメタル式EX温度センサ　　精度±1℃
湿度計：バイマテリアル式EX湿度センサ　　精度±3%RH
（EMPEX TM-6251 取扱説明書より）

物差しの例

ステンレス直尺15cm　　誤差±0.15mm → ¥390
ステンレスノギス150mm　　誤差±0.05mm → ¥6,810
デジタルノギス150mm　　誤差±0.03mm → ¥9,910
（シンワ測定株式会社HPより；2024/8/24現在）

もちろん，誤差の小さい（精度の高い）測定がよい
しかし，精度を上げると手間とコストがかかる

人間の特性の測定

人間のもつ心理的特性（能力，性格，態度etc.）に対して，数値を割り当てる

構成概念（construct）

「～力」「～性」などのラベルはあるが，学問的な仮説や社会的要請に基づくものがほとんど。実在するかどうかはわからない抽象化された概念。

測定したいものを直接測定することができない

「質問紙」「テスト」などを使い，目に見える行動や反応を記録することによって間接的に測定する

繰り返し測定は難しい

測定対象は複数（多数の受検者）

心理測定学とテスト理論

心理測定学（psychometrics）

人間の様々な心理学的特性を測定するための方法論に関する学問体系

教育分野では教育測定学 (educational measurement)

心理測定学の専門家を，サイコメトリシャン (psychometrician)という

テスト理論（test theory）

心理測定学の理論のひとつ

テストを使った測定において，その性能を定量的・統計的に評価するための基礎を与える理論

古典的テスト理論 (classical test theory: CTT)

項目反応理論 (item response theory: IRT)

テストによる測定の評価の観点

T. L. Kelley

妥当性は，テストが測定すべきものを測定しているかという問題であり，信頼性は，テストが測定するものを正確に測定しているかという問題である。(p. 14)

テスト得点の精度は？

誤差（本来あるべき得点からのズレ）の小ささ

→ 信頼性 (reliability)

テストが，測りたいものをちゃんと測れているか？

テストの得点が，測りたい特性を的確に反映しているか？

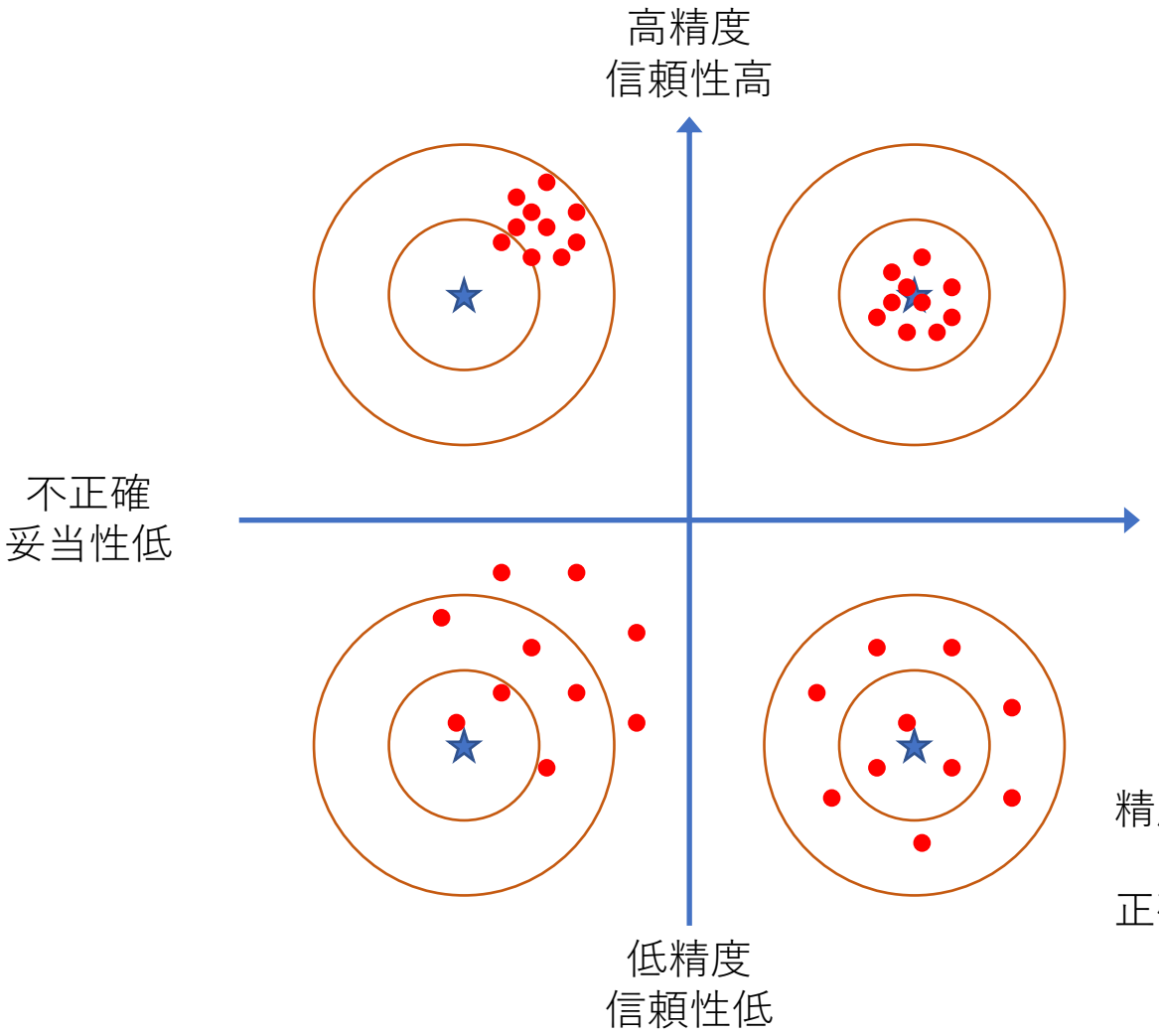
テストの得点をどう解釈できるか？

→ 妥当性 (validity)

※測定対象が「見えない」が故の，テストに固有の問題

Kelley, T. L. (1927). *Interpretation of educational measurements*. Macmillan.

テストの信頼性と妥当性



精度（precision）
＝偶然誤差の小ささ
正確度（accuracy）
＝系統誤差の小ささ

テスト作成のゴール

「同じ能力＝同じ得点」となるような
測定器具（＝テスト）
尺度（＝モノサシ）
を作ること

テストの品質（性能）
「信頼性」と「妥当性」
その他に，「公平性」「比較可能性」「有用性」etc.

テスト作成のステップ

1. テストの目的・用途を決める
2. テストのフレームワークを作る
   a. 構成概念の定義
   b. 行動指標の同定・収集
   c. テストおよび項目の仕様を決める
3. テスト項目の作成・レビュー
4. 予備実査
   a. データの収集
   b. データの分析（信頼性の確認／尺度構成・等化）
5. テストの編集
6. スコア・レポート，解釈基準やマニュアル等の作成

すべての段階において，妥当性を高める努力が必要

加藤健太郎 (2023). テストの作り方－能力・学力テストを中心として－ 繁桝算男（編） 心理・教育・人事のためのテスト学入門 誠信書房

テストの種類

内容
- 能力，学力，適性，性格，興味，態度etc.

実施形態
- 個人検査vs集団検査
- PBT（紙で実施），CBT（コンピュータで実施），器具の使用（実技）etc.

評価の仕方
- 規準集団準拠テスト（norm-referenced test）＝規準となる集団と比較した相対評価
- 内容基準準拠テスト（criterion-referenced test）＝予め設定された基準と比較した絶対評価

評価の文脈
- 包括的評価（summative）
- 形成的評価（formative）

結果の利用
- 結果の評価単位（個人／集団）
- 結果の利用者（個人／組織）

渡部洋（編著） (1993). 心理検査法入門 福村出版

テスト開発者の責務

テストは，その目的や用途に照らして有用な情報を提供できるようにデザインされるべき

1つのテストを形にし，実施するまでには，様々な課題や制約がある

実施形態・環境，テスト時間，規模，開発・実施・結果返却までのスケジュール，予算・人員，問題の形式，採点体制etc.

テストの目的・用途に照らして，やるべきこと／できないことの優先順位付けをする必要がある

テストの根本的な機能は，あくまで測定

すべてのステークホルダーに対して，きちんと説明責任が果たせること

古典的テスト理論モデルと
テストの信頼性

テスト得点は変動する

変動の要因は何か？

テスト得点について・・・

個人間　測定したい特性における個人差が忠実に反映されていることが望ましい

個人内　（測定したい特性が一定なら）一貫していることが望ましい

誤差（error）

テスト得点に含まれる，測定したい特性に無関係な成分

誤差成分の大きさが，測定の精度を決める

テストの実施過程で生じる誤差

1. 採点に関わる誤差
   採点者の評価のばらつき，疲労・慣れによる変化
2. 受検者の解答プロセスで生じる誤差
   当て推量，うっかりミス，反応バイアス
3. 受検者に内在する誤差
   テストの「学習」効果，受検者の健康状態，集中力・疲労
4. テスト項目の抽出における誤差

野口裕之・大隅敦子 (2014). テスティングの基礎理論 研究社

誤差の分類

系統誤差（systematic error／bias）

個人（or集団）の測定値に一貫して含まれる誤差

測定回数を増やしても消えない

例：反応歪曲，事前知識の有無，カリキュラムの違い，文化差，性差etc.

偶然誤差（random error）

不確実な要因によって生じる誤差

測定回数を増やすことによって相殺される

例：当て推量，うっかりミス，個人の「調子」，採点ミスetc.

仮想実験

一人の生徒がAとBの2つのテストを繰り返し受験したところ，得点は以下のようになった：

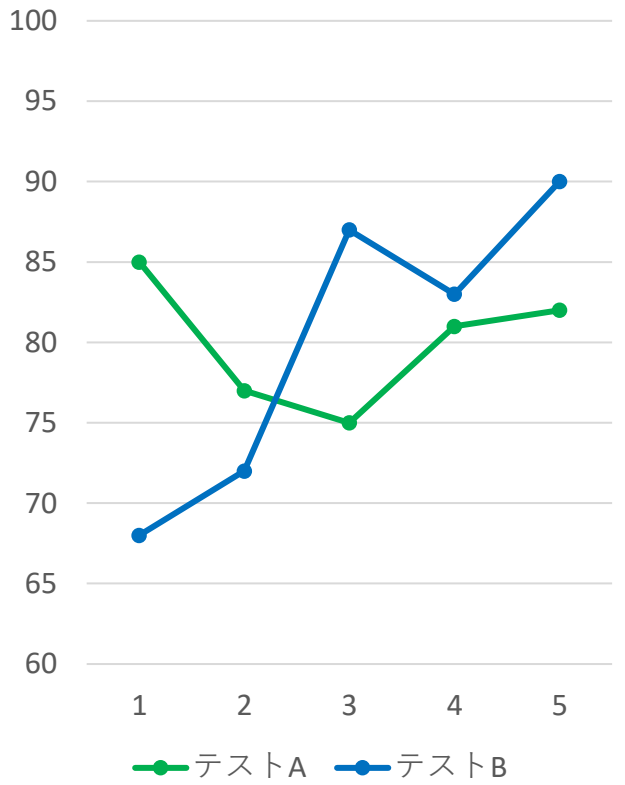
テストA: 85, 77, 75, 81, 82

テストB: 68, 72, 87, 83, 90

どちらのテストも平均は80点

標準偏差は3.58（テストA）と8.56（テストB）

どちらが「よい」テスト？



CTTモデル(1)：個人内モデル

任意の個人（iと表す）についてのモデル： $X_i = t_i + E_i$

X_i：テスト得点（観測される；E_iが確率変数なので，X_iも確率変数）

t_i：個人iの真値（観測できない；個人内では常に定数）

E_i：偶然誤差（観測できない；測定の度に変動する → 確率変数と見なす）

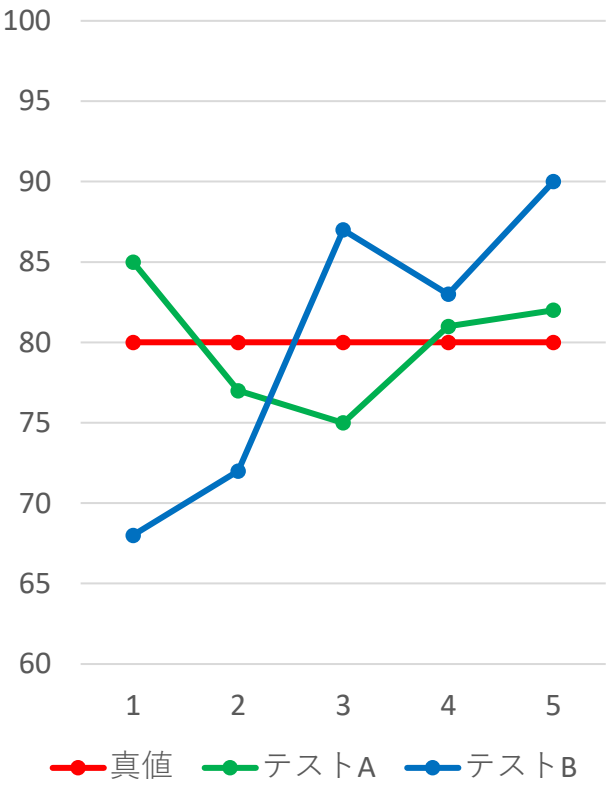
誤差E_iの分散をσ_E^2と表し，これを誤差分散と呼ぶ

測定精度＝偶然誤差の小ささ＝σ_E^2の小ささ

※誤差を確率変数と見なすということは，「個人iについて（無限回の）繰り返し測定を行うことができるとしたらどうなる」という仮説的な状況を想定しているということ

真値と偶然誤差

	テストA			テストB		
	X =	t	+ E	X =	t	+ E
	85 =	80	+ 5	68 =	80	- 12
	77 =	80	- 3	72 =	80	- 8
	75 =	80	- 5	87 =	80	+ 7
	81 =	80	+ 1	83 =	80	+ 3
	82 =	80	+ 2	90 =	80	+ 10
平均	80	80	0	80	80	0
標準偏差	3.58	0	3.58	8.56	0	8.56
分散	12.80	0	12.80	73.20	0	73.20



CTTモデル(2)：個人間モデル

被験者の母集団を考える

個人iは母集団からの無作為標本（と見なす）

各個人については，個人内モデルが成り立っているとする

母集団から抽出される個人iについてのモデル： $X_i = T_i + E_i$

T_i：個人iの真値（観測できない；誰がサンプルされるかによって変動するので，確率変数と見なす）→ 真値T_iの分散をσ_T^2と表す

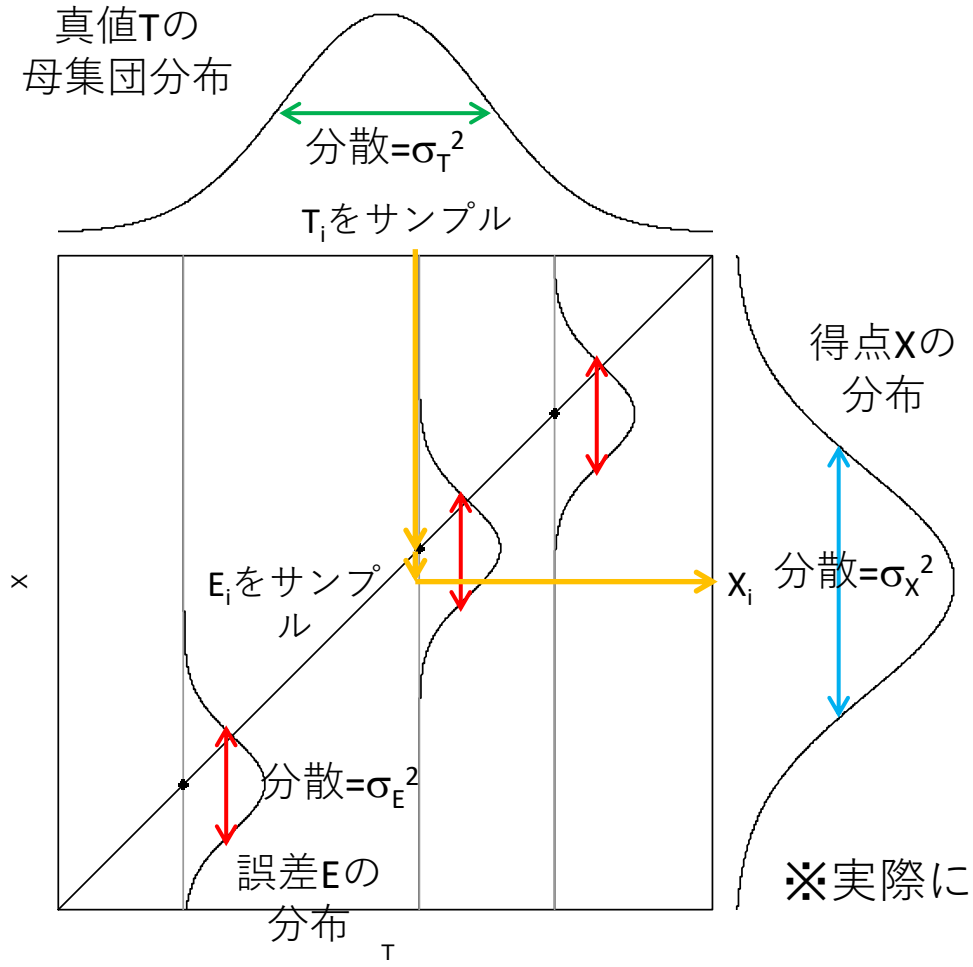
X_iとE_iは個人内モデルと同じ

真値どうし，誤差どうし，真値と誤差が互いに無相関であると仮定すると，テスト得点X_iの分散σ_X^2が

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

と表せる

CTTモデル：全体の状況



CTTモデル

$$X = T + E$$

テスト得点の分散は，真値の分散と誤差の分散の和に分割できる

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

偶然誤差の「大きさ」＝誤差分散σ_E^2

※実際に観測できるのはテスト得点Xのみ

信頼性の定義

分散の分割式：$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$
→ σ_E^2が小さい方が望ましい

テスト得点Xの信頼性係数（reliability coefficient）：

$$\rho_X = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$$

※真値による分散説明率（proportion of variance explained: PV）＝「テスト得点の変動のうち，何割が真値の変動を反映しているのか？」

$0 \leq \rho_X \leq 1$
$\rho_X = 0$なら，得点の変動は偶然誤差のみによる（真値の違いをまったく反映しない）
$\rho_X = 1$なら，得点の変動は真値の違いのみによる（誤差がまったくない；得点は常に真値と一致）

信頼性係数の推定

知りたいのは誤差分散（σ_E^2）の大きさ

テストデータから直接推定するのは難しい → どうする？

再テスト法

同じ（あるいは等価な）テストを2回実施し，得点間の相関係数を求める

内的一貫性にもとづく方法

折半法：1つのテストを半分に分割し，部分テストの得点間の相関係数を求めて補正する

Cronbachのα係数

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum s_j^2}{s_x^2}\right)$$

・・・項目j の得点の標本分散（s_j^2）およびテスト得点の標本分散（s_x^2）の推定値から計算。 Jは項目数。

McDonaldのω係数

$$\omega = \frac{\left(\sum \hat{\lambda}_j\right)^2}{\left(\sum \hat{\lambda}_j\right)^2 + \sum \hat{\sigma}_{E_j}^2}$$

・・・ 1因子の因子分析を行い，各項目の因子負荷（λ_j）および独自性（$\sigma_{E_j}^2$）の推定値から計算

信頼性はどれくらいあればよい？

絶対的な基準はない

0.8程度を目安とすることが多いが，テストの利用目的や使い方による

受検者の集団によっても異なる（※信頼性はテスト固有の性能ではない）

測定の標準誤差も合わせて参照すべき（誤差の標準偏差σ_Eの推定値で，$\hat{\sigma}_E = s_x\sqrt{1-\hat{\rho}_X}$で計算される）

△信頼性が低いテスト

テストの利用目的が達成できない（テスト得点にもとづく意思決定を誤る）可能性

「効果」や「関連」を確認しづらい

（得点の出方がおかしいときに）原因を特定しづらい

例：SAT

Test Characteristics of the SAT®: Reliability, Difficulty Levels, Completion Rates

January 2012–December 2012

	Critical Reading (67 Questions)	Mathematics (54 Questions)	Writing MC (49 Questions)	Writing Composite* (with Essay)
Reliability coefficient	.91-.92	.91-.93	.89-.91	.89-.91
Standard error of measurement (SEM)	30–32	28–32	3.4–3.5	31–35
Standard error of difference (SED)	42–45	40–45	5	44–49
Difficulty (average percent correct)	.55–.61	.55–.62	.58–.66	

*Essay reliability statistics, which are the proportion of essays that are scored the same or within one point, range between .99 and .100.

※旧SAT Reasoning Test
Critical Reading， Mathematics， Writingの3教科， それぞれ得点のレンジは200～800点

https://secure-media.collegeboard.org/digitalServices/pdf/research/Test-Characteristics-of-SAT-2013.pdf

信頼性を高めるには

項目数を増やす
- 項目数大 → 信頼性大
- 既存の項目やテスト得点との相関がほとんどゼロの場合は効果小（負の相関なら逆効果）

相関を高める
- 項目どうしに正の相関
- 個々の項目とテスト得点の間に正の相関（IT相関，項目識別力）

テストの内容
- 出題領域が多様な要素を含むほど，信頼性は下がる傾向

項目の形式
- 客観式（多肢選択，短答など）は，そうでない形式（記述式，論述式など）よりも信頼性が高くなる傾向

集団の異質性（heterogeneity）
- 対象とする受験者の集団が，真値に関して異質な（i.e., 真値の分散（σ_T^2）が大きい）ほど，信頼性は高くなる

受検者の状態
- 動機付け，集中力，疲労，テスト慣れ，カンニングなど，測定したい特性以外の要因が得点に影響すると，信頼性は下がる
- テスト環境や手続きが標準化されていることが望ましい（制限時間，時間割，実施および採点の方法など）

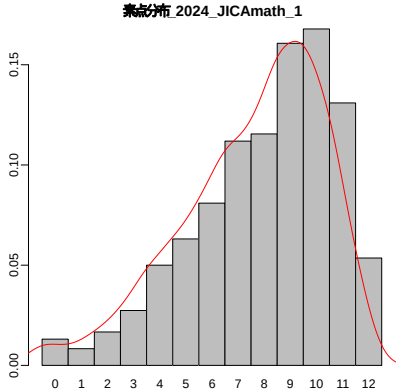
CTTにもとづくJ-MAGPAの分析結果(テスト分析)

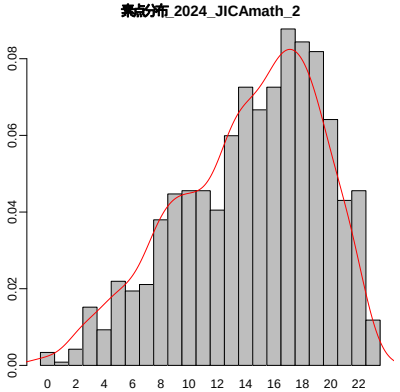
対象学年	受検者数	問題数	平均正答率	標準偏差	最低	Q1	中央値	Q3	最高	α係数	測定の標準誤差	0点人数	満点人数
1	840	12	0.670	0.224	0.000	0.500	0.750	0.833	1.000	0.727	0.117	11	45
2	1185	23	0.635	0.216	0.000	0.478	0.652	0.783	1.000	0.851	0.083	4	14
3	1266	24	0.592	0.208	0.000	0.458	0.625	0.750	1.000	0.838	0.084	4	4
4	1240	26	0.483	0.227	0.000	0.308	0.500	0.654	0.962	0.880	0.079	10	0
5	1257	32	0.442	0.184	0.000	0.313	0.438	0.594	0.875	0.856	0.070	2	0
6	1087	33	0.516	0.152	0.000	0.424	0.515	0.606	0.909	0.792	0.069	3	0

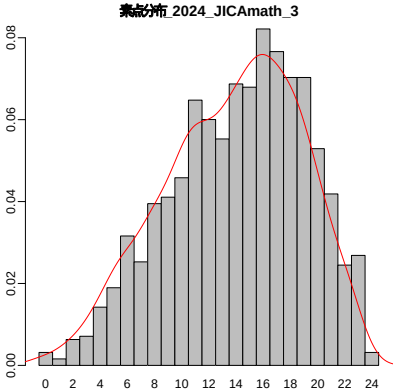
主なチェックポイント

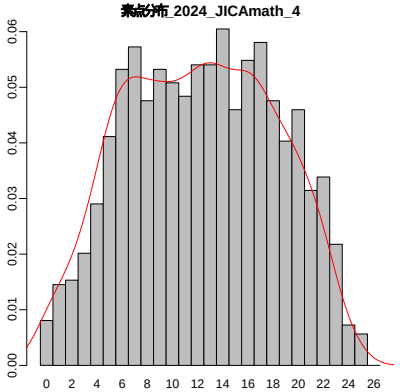
分布の位置・散布度・形状

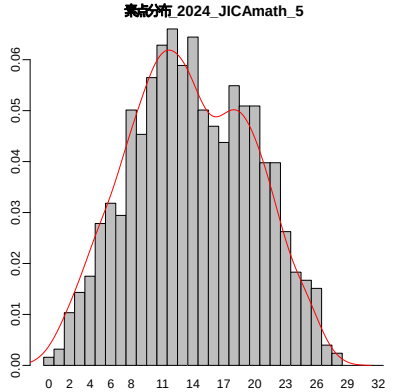
測定精度（信頼性・測定誤差）

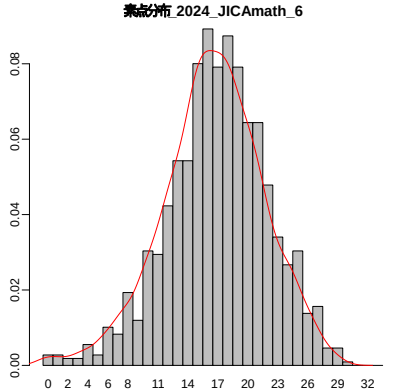












CTTにもとづくJ-MAGPAの分析結果(項目分析)

対象学年	項目ID	受検者数	有効解答数	無効解答率	未到達率	平均正答率	標準偏差	IT相関	除外IT相関	除外α係数
2	N111AM1G2	1185	1073	0.095	0.000	0.808	0.394	0.445	0.378	0.847
2	N113M1G2	1185	1173	0.010	0.000	0.893	0.309	0.438	0.385	0.847
2	N113M2G2	1185	1165	0.017	0.000	0.642	0.480	0.483	0.403	0.846
2	N113M3G2	1185	1160	0.021	0.000	0.414	0.493	0.511	0.432	0.845
2	N131M1G1	1185	1183	0.002	0.000	0.921	0.270	0.340	0.290	0.850
2	N131M2G1	1185	1183	0.002	0.000	0.881	0.324	0.417	0.361	0.848
2	N131M3G1	1185	1184	0.001	0.000	0.669	0.471	0.599	0.532	0.841
2	N131M4G1	1185	1182	0.003	0.000	0.684	0.465	0.614	0.550	0.840
2	N131M1G2	1185	1180	0.004	0.000	0.585	0.493	0.588	0.517	0.841
2	N131M2G2	1185	1179	0.005	0.000	0.707	0.455	0.340	0.255	0.851
2	N131M3G2	1185	1178	0.006	0.000	0.723	0.448	0.404	0.324	0.849
2	N131M4G2	1185	1175	0.008	0.000	0.537	0.499	0.547	0.470	0.843
2	N133M1G2	1185	1145	0.034	0.000	0.348	0.476	0.493	0.415	0.846
2	N133M2G2	1185	1132	0.045	0.002	0.319	0.466	0.462	0.383	0.847
2	N141M1G1	1185	1094	0.077	0.003	0.566	0.496	0.523	0.444	0.844
2	N141M2G1	1185	1063	0.103	0.010	0.569	0.495	0.631	0.565	0.839
2	M111AM1G2	1185	1097	0.074	0.011	0.872	0.335	0.379	0.319	0.849
2	M212M1G2	1185	1083	0.086	0.013	0.505	0.500	0.408	0.319	0.850
2	M214M1G2	1185	988	0.166	0.031	0.432	0.496	0.527	0.449	0.844
2	G111M1G1	1185	1062	0.104	0.040	0.797	0.403	0.581	0.524	0.842
2	G211M1G2	1185	994	0.161	0.056	0.324	0.468	0.410	0.327	0.849
2	S112M1G2	1185	1022	0.138	0.062	0.597	0.491	0.468	0.386	0.847
2	A111M1G2	1185	1077	0.091	0.072	0.806	0.396	0.508	0.445	0.845

前後と比べて無解答が多い項目 → 内容に問題？

時間切れ？

主なチェックポイント

無効解答率・未到達率

正答率

IT相関（項目識別力） ※ 測定精度に関係

→ 数値に問題がある項目は，実施状況や内容とも突き合わせて使用・修正・廃棄を判断

※教科テストの場合，単元を履修済みかどうかで差が出るので注意

CTTにもとづくJ-MAGPAの分析結果(項目分析)

対象学年	項目ID	受検者数	有効解答数	無効解答率	未到達率	平均正答率	標準偏差	IT相関	除外IT相関	除外α係数
5	N122M1G5	1257	993	0.210	0.000	0.339	0.474	0.522	0.460	0.850
5	N122M2G5	1257	1104	0.122	0.000	0.456	0.498	0.546	0.483	0.849
5	N123M1G5	1257	738	0.413	0.000	0.050	0.218	0.141	0.104	0.857
5	N321M1G5	1257	699	0.444	0.000	0.059	0.235	0.159	0.120	0.857
5	N131M1G2	1257	1253	0.003	0.000	0.883	0.321	0.336	0.286	0.854
5	N131M2G2	1257	1257	0.000	0.000	0.876	0.330	0.197	0.143	0.857
5	N131M3G2	1257	1257	0.000	0.000	0.901	0.299	0.254	0.205	0.856
5	N131M4G2	1257	1249	0.006	0.000	0.815	0.389	0.384	0.326	0.853
5	N131M1G4	1257	1242	0.012	0.000	0.751	0.433	0.502	0.444	0.850
5	N131M2G4	1257	1216	0.033	0.000	0.426	0.495	0.489	0.421	0.851
5	N135M1G4	1257	1235	0.018	0.000	0.717	0.451	0.443	0.358	0.852
5	N133M1G4	1257	1199	0.046	0.000	0.586	0.493	0.557	0.495	0.848
5	N133M1G5	1257	1148	0.087	0.000	0.308	0.462	0.552	0.494	0.849
5	N133M2G4	1257	1117	0.111	0.000	0.468	0.499	0.595	0.536	0.847
5	N133M2G5	1257	1073	0.146	0.000	0.309	0.462	0.561	0.503	0.848
5	N137M1G4	1257	1162	0.076	0.000	0.111	0.315	0.253	0.202	0.856
5	N221M1G4	1257	1124	0.106	0.000	0.336	0.472	0.419	0.350	0.853
5	N221M1G5	1257	1088	0.134	0.000	0.029	0.167	0.098	0.069	0.857
5	N33M1G5	1257	1141	0.092	0.000	0.600	0.490	0.574	0.515	0.848
5	N212M1G5	1257	1073	0.146	0.000	0.265	0.441	0.396	0.331	0.853
5	N214M1G5	1257	1102	0.123	0.002	0.103	0.305	0.188	0.137	0.857
5	N141M1G5	1257	1081	0.140	0.006	0.476	0.500	0.588	0.528	0.847
5	N141M2G5	1257	1077	0.143	0.011	0.267	0.443	0.514	0.455	0.850
5	M112AM1G4	1257	1127	0.103	0.012	0.614	0.487	0.440	0.370	0.852
5	M113AM1G5	1257	1082	0.139	0.018	0.257	0.437	0.302	0.233	0.856
5	M122M1G5	1257	1097	0.127	0.021	0.286	0.452	0.472	0.409	0.851
5	G112AM1G5	1257	1150	0.085	0.024	0.260	0.439	0.232	0.160	0.858
5	G212M1G5	1257	1134	0.098	0.025	0.412	0.492	0.490	0.423	0.851
5	S113E1G4	1257	1110	0.117	0.027	0.477	0.500	0.482	0.413	0.851
5	S211M1G5	1257	1064	0.154	0.033	0.473	0.499	0.475	0.405	0.851
5	A112M1G4	1257	1131	0.100	0.041	0.511	0.500	0.468	0.398	0.851
5	A323M1G5	1257	1170	0.069	0.054	0.736	0.441	0.510	0.452	0.850

前後と比べて極端に無解答が多い→内容に問題？

正答率が極端に低い

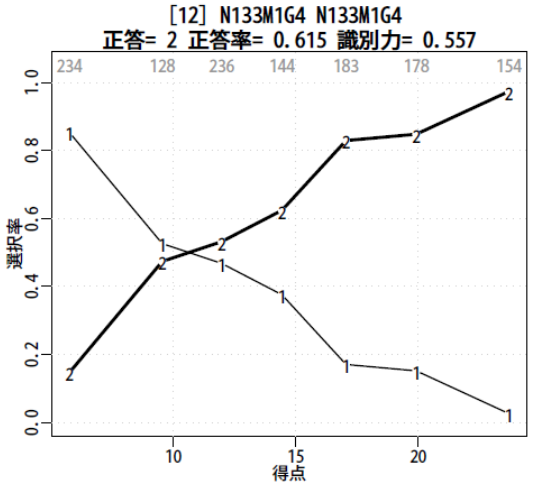
項目識別力が低い

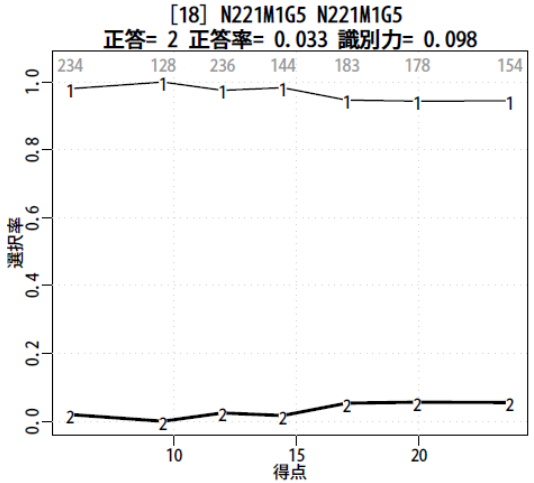
除外すると信頼性が高くなる

トレースライン

項目分析の結果を視覚化

正答のラインが右上がりになるのが望ましい





妥当性とその検証

「妥当性」観の変遷

妥当性の検証（validation）

テスト結果の解釈や使用がどの程度尤もらしいか（plausibility）を評価すること，およびそのための証拠（evidence）を集積すること

テスト作成・運用上必要不可欠な作業

<u>妥当性の概念は時代とともに変化してきた</u>

～1955

基準妥当性

内容妥当性

1955～1989

構成概念妥当性

1989～

構成概念妥当性としての統合的な評価

Kane, M. T. (2006). Validation. In R. B. Brennan (Ed.), *Educational measurement (4th ed.)* (pp. 17-64). American Council on Education/Praeger.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* (pp. 13-103). American Council on Education/Macmillan.
村山航 (2012). 妥当性概念の歴史的変遷と心理測定学的観点からの考察 教育心理学年報 , 51, 118-130.

「古典的な」妥当性観

基準妥当性

外的基準との関連性（併存的妥当性・予測的妥当性・増分妥当性）

「妥当性係数」

内容妥当性

項目・テストの内容の尤もらしさ（構成概念との関連性，測定領域の網羅性・代表性）

構成概念妥当性

理論的に予想される他の構成概念との関連性が見られるか（収束的妥当性，弁別的妥当性）

広義の「構成概念妥当性」への統合

S. J. Messick

妥当性の諸概念を，広義の構成概念妥当性として統一的に捉えるべきである。
そのような意味での構成概念妥当性は，「主張」に対する「証拠」の蓄積（の程度）である。

現代の妥当性理論は，Messickの妥当性概念と，これを踏まえたKaneの妥当性検証のアプローチが主流

平井洋子 (2006). 測定の妥当性から見た尺度構成－得点の解釈を保証できますか 吉田寿夫（編著）『心理学研究法の新しいかたち』 誠信書房 pp. 21-49
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* (pp. 13-103). American Council on Education/Macmillan.
清水裕子 (2005). 測定における妥当性の理解のために―言語テストの基本概念として― 立命館言語文化研究, 16, 241-254.

妥当性検証のロジック：論証ベースアプローチ

テスト得点の解釈に基づく「主張」（claim）

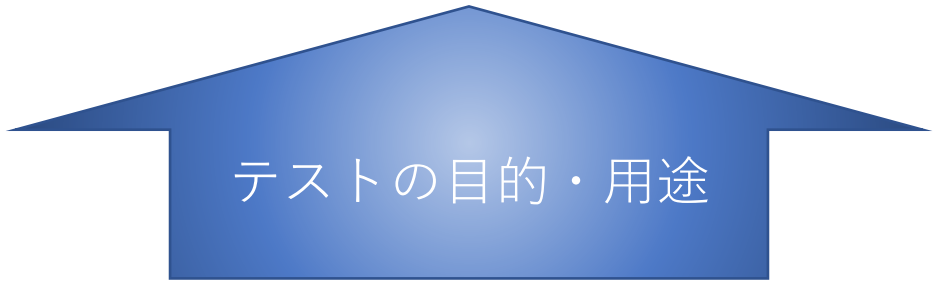
「このテストは動機付けという特性を測定している」

「○×点以上であれば，生活に必要な英語力が十分にある」

「△※点以上であれば，臨床的介入が必要である」

「このテストを用いることによって，より適切に資格のある人を選抜することができる」

・・・



妥当性検証のロジック：論証ベースアプローチ

その主張を支える理論的な根拠は何か？

- 構成概念の背景にある理論や仮説の合理性

その主張を裏付ける証拠は？

- 目的に応じて，様々な形で証拠を集める

妥当性の検証（validation）は「主張」を裏付ける「証拠集め」の繰り返し

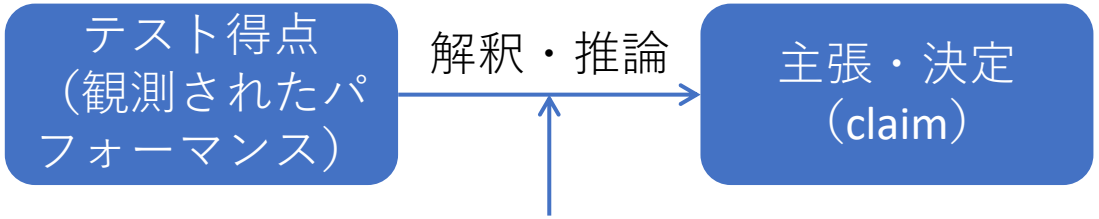
- テストの作成段階に始まり，運用開始後も継続的に検証を行っていく
- テストの目的・用途，フレームワーク（背景理論や仮説）がないと評価できない

構成概念妥当性（再定義）

- テスト得点の解釈を裏付ける証拠の蓄積が構成概念妥当性
  - 理論的な説明に合致する情報ほど証拠として強い
  - 対立仮説も評価・・・他の説明可能性が排除されることにより，提案している解釈が強化される（反証主義）
  - 旧来の構成概念妥当性の検証（テストの内的，外的構造，収束的・弁別的妥当性）のみにとどまらない

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.

構成概念妥当性を構成する側面



構成概念妥当性の6つの側面

- 内容的側面（content）
- 本質的側面（substantive）
- 構造的側面（structural）
- 一般化の側面（generalization）
- 外的側面（external）
- 結果的側面（consequential）

各側面において，構成概念妥当性を強化する証拠は何か？

各側面は個別に「～妥当性」を構成するのではなく，相互に関連しながら「テスト得点の解釈」を裏付けるために必要な要素として統合的に機能する

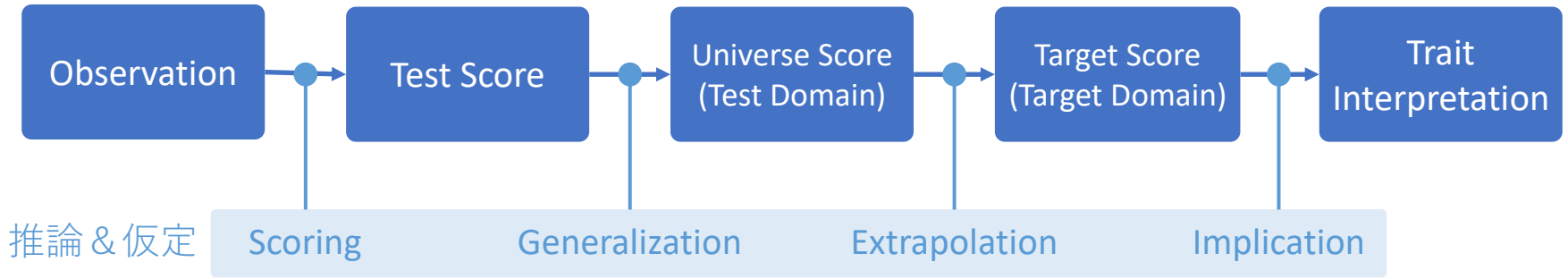
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.

構成概念妥当性の６つの側面

側面	説明	主な検証方法
内容的側面	従来の「内容妥当性」に相当	作問時，作問後の専門家によるテスト項目のレビュー 構成概念やテスト領域の定義に関する専門家のレビュー
本質的側面	その項目orテストが，意図している思考過程やスキルの使用，心理的反応を喚起しているか	観察，思考発話（think-aloud）法，反応時間，解答プロセスの認知／数理モデリングなど
構造的側面	従来の狭義の「構成概念妥当性」に相当 テストの内的構造や採点の適切さも含む	テストデータの解析（因子分析・構造方程式モデリング等による仮説検証） 採点基準の確認
一般化の側面	個人or集団のテスト結果が，他の状況にどの程度一般化できるか？	CTTやIRTにもとづく分析 交差妥当化，他の行動指標との突き合わせ
外的側面	従来の「基準妥当性」「構成概念妥当性」に相当	テストデータの解析（因子分析・構造方程式モデル・回帰分析等による仮説検証）
結果的側面	テスト得点の使用によって，どのような（意図した／していなかった，ポジティブ／ネガティブな）結果がもたらされたか？	調査・プログラム評価（定性・定量）

「解釈・推論」と妥当性検証の枠組み

Interpretive arguments for trait interpretations (Kane, 2006)



→ これらの推論の適切さを合理的説明と証拠にもとづく評価を記述したものがvalidity arguments

Kane, M. T. (2006). Validation. In Brennan, R. L. (Ed.), *Educational measurement* (4th ed., pp. 17-64). ACE/Praeger.

妥当性を高めるには

テストを「意図して設計し，作る」ことが大切

- すべての基本となるのは「テストの目的・用途」＆「テストのフレームワーク」→ ここがないテストは妥当性の評価すらできない
- 妥当性を「確認する」「確立する」のではなく，「高める努力をする」「どれだけの証拠があるか」

重要な妥当性の側面・推論に関する証拠を収集する

- 関連する様々な情報（量的・質的両面）を継続して蓄積していく
- 証拠の質を保証するための，適切な制作手続きや検証計画も大切

J-MAGPAの妥当性検証の観点

目的

算数学び隊派遣国の子ども達の算数に係る学習到達度を把握・分析し、その改善を図るとともに、算数学び隊の配属先における子ども達の教育指導の充実や学習状況の改善等に役立てること

具体的な用途は？

評価の文脈（形成的or総括的）

評価の仕方（基準集団準拠or内容基準準拠）

結果の利用（個人or集団）

→ こうした分類のどこに落ちるかを明確にすると，今後のテストの設計・開発方針や性能検証の観点を定めやすい

J-MAGPAの妥当性検証の観点

内容的側面

GPF準拠（Global Proficiency Framework: Reading and Mathematics | EducationLinks (edu-links.org)）

→ 単元（construct）および学年に対応したレベルが詳細に記述されている（test domainが明確）

→ 関連性・網羅性の判断が非常に行いやすい

Domain	Construct		Subconstruct		Grade								
					1	2	3	4	5	6	7	8	9
N Number and operations	N1	Whole numbers	N1.1	Identify and count in whole numbers, and identify their relative magnitude	x	x	x	x	x	x	a	a	a
			N1.2	Represent whole numbers in equivalent ways	x	x	x	x	x	x	a	a	a
			N1.3	Solve operations using whole numbers	x	x	x	x	x	x	see integers		
			N1.4	Solve real-world problems involving whole numbers	x	x	x	x	x	x	see integers		
	N2	Fractions	N2.1	Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude			x	x	x	x	x	a	a
			N2.2	Solve operations using fractions				x	x	x	x	a	a
			N2.3	Solve real-world problems involving fractions				x	x	x	x	a	a
	N3	Decimals	N3.1	Identify and represent decimals using objects, pictures, and symbols, and identify relative magnitude					x	x	x	a	a
			N3.2	Represent decimals in equivalent ways (including fractions and percentages)					x	x	x	x	a
			N3.3	Solve operations using decimals					x	x	x	x	a
			N3.4	Solve real-world problems involving decimals						x	x	x	a
	N4	Integers	N4.1	Identify and represent integers using objects, pictures, or symbols, and identify relative magnitude							x	a	a
			N4.2	Solve operations using integers							x	x	a
			N4.3	Solve real-world problems involving integers							x	x	a
	N5	Exponents and roots	N5.1	Identify and represent quantities using exponents and roots, and identify the relative magnitude							x	x	x
			N5.2	Solve operations involving exponents and roots								x	x
	N6	Operations across number	N6.1	Solve operations involving integers, fractions, decimals, percentages, and exponents								x	x

Global Proficiency Framework for Mathematics (https://www.edu-links.org/sites/default/files/media/file/GPF-Math-Final.pdf)

J-MAGPAの妥当性検証の観点

構造的側面

単元ごとの到達度を測定したいのか（→単元ごとの得点），すべての単元を含めた総合的な到達度を測定したいのか（→全単元を通じた得点）？

対応した内的構造があるか？（全体で一次元？領域ごとにまとまる？）（←CTT分析の結果はほぼ一次元）

一般化の側面

個人差を測定するテストとしての性能に大きな問題はない（←CTT分析の結果）

「到達度」に関して

単元ごとの到達度に関する推論（test domainへの一般化やその先）は弱い（単元ごとの項目数が少ない）

総合的な到達度 → おそらくperformance setting（何点取れればproficientか？何点の子どもは何ができるのか）が必要

グローバルに実施するという観点から，項目特異機能（differential item functioning: DIF）のチェックは必要

結果的側面

J-MAGPAを利用することによって，当初の目的が達成できたか？

どんなポジティブ・ネガティブな影響があったか？

参考文献

池田央 (1992). テストの科学 日本文化科学社
池田央 (1994). 現代テスト理論 朝倉書店
日本テスト学会（編） (2007). テスト・スタンダードー日本のテストの将来に向けて 金子書房
日本テスト学会（編） (2010). 見直そう，テストを支える基本の技術と教育 金子書房
野口裕之・大隅敦子 (2014). テスティングの基礎理論 研究社
繁桝算男（編）(2023). 心理・教育・人事のためのテスト学入門 誠信書房

AERA/APA/NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
Crocker, L., & Algina, J. (1987/2006). *Introduction to classical and modern test theory*. Holt Reinhart Winston/Wadsworth.

（その他，本資料の各所で挙げた文献をご覧ください）

ご清聴ありがとうございました

EOF