話すこと(やりとり)
テスト用タスクの予備調査報告
(発表後にいただいたコメントに基づき若干修正を加えてあります)

小泉　利恵
順天堂大学

金子　恵美子
会津大学

印南　洋
中央大学

長沼　君主
東海大学

予備調査の目的

- 以下の2点を確認する
  - CEFR-Jに基づいて作成した「話すこと（やりとり）」のテストタスクが、そのレベルに応じた難易度になっているか。
  - タスクが測る能力が意図通りか。

予備調査対象タスク

- PreA1, A1.1, A1.2, A1.3, A2.1の5レベル
- 各2タスク（CEFR-Jの記述子通り）
- 合計　10タスク

CEFR-J　話すこと（やりとり）

レベル	PreA1	A1.1	A1.2	A1.3	A2.1
やりとり	基礎的な語句を使って，「助けて！」や「～が欲しい」などの自分の要求を伝えることができる。また，必要があれば，欲しいものを指さししながら自分の意思を伝えることが出来る。	なじみのある定型表現を使って、時間・日にち・場所について質問したり、質問に答えたりすることができる。	基本的な語や言い回しを使って日常のやりとり(何ができるかできないかや色についてのやりとりなど)，において単純に応答することができる。	趣味、部活動などのなじみのあるトピックに関して、はっきりと話されれば、簡単な質疑応答をすることができる。	順序を表す表現であるfirst, then, nextなどのつなぎ言葉や「右に曲がって」や「まっすぐ行って」などの基本的な表現を使って，単純な道案内をすることができる。
	一般的な定型の日常の挨拶や季節の挨拶をしたり、そうした挨拶に応答したりすることができる。	家族、日課、趣味などの個人的なトピックについて、（必ずしも正確ではないが）なじみのある表現や基礎的な文を使って、質問したり、質問に答えたりすることができる。	スポーツや食べ物などの好き嫌いなどのとてもなじみのあるトピックに関して，はっきり話されれば，限られたレパートリーを使って，簡単な意見交換をすることができる。	基本的な語や言い回しを使って，人を誘ったり，誘いを受けたり，断ったりすることができる。	補助となる絵やものを用いて，基本的な情報を伝え，また，簡単な意見交換をすることができる。

テスト形式

テスト実施環境の統一化を図る

- 試験官（教員）一人と受験者（学生）の対面
- 各タスク回答のための制限時間（1分）
- タスクを理解するのに時間が必要なものには一定の準備時間
- タスクの指示はカードで提示（日本語）

予備調査のための試行調査

- A1以上の口頭運用能力があると思われる学生9名
- 10タスクのうち、平均3タスクを実施
- ビデオで録画

修正点

- 発話時間を1分半～3分へ
- 指示カードの簡略化

評価基準合わせ

ベンチマークとして利用したもの

1. DVD "Spoken performances illustrating the 6 levels of the Common European Framework of Reference for Languages"

●http://www.ciep.fr/en/books-and-cd-roms-dealing-with-assessment-and-certifications/dvd-spoken-performances-illustrating-the-6-levels-of-the-common-european-framework-of-reference-1

● *It is the product of a seminar organized in 2008 by CiEP with the assistance of the Language Policy Division of the Council of Europe and with the participation of Cambridge ESOL, the Cervantes Institute, the Eurocentres Foundation, the Goethe Institute, and the University of Perugia Centre for Assessment and Language Certification.*

評価基準合わせ

2. Cambridge Assessmentのサイトにある Examples of Speaking tests

   Cambridge English: Key (KET)の

   動画（A2）とコメント

●http://www.cambridgeenglish.org/research-and-validation/fitness-for-purpose/

評価基準合わせ

3. CEFR準拠のインタビュー実施のためのサンプル動画
   – 必ずしも、受験者がそのレベルではない

- 試行調査に参加した学生の発話を第一著者、第二著者で評価し、評価基準を統一するための打ち合わせ

評価

- インタビュー実施と同時に試験官が3段階評価
  - 評点3：期待以上のパフォーマンスでタスクを遂行
    - 一貫してeffortlessnessである
    - 自分から話そうという前向きな姿勢に加えて、流暢さ、正確さなど、もう一点特に強い面がある
  - 評点2：そのレベルで期待されるパフォーマンスでタスクを遂行
  - 評点1： CEFR-Jレベルで期待されるパフォーマンスに未達
- 予備調査の約20%は試験官2人によるダブルレイティング

予備調査実施

- 2017年12月から2018年1月に実施
- 参加者
  - 順天堂大学：58名（試行調査9名を含む）
  - 会津大学：18名
  - 合計：66名

Can-do自己評価

- インタビュー終了後にアンケートを実施
- CEFR-Jの「～できる」という記述子、Pre-A1からB1.2まで2項目ずつ、全16項目
    1. ほとんどできない
    2. あまりできない
    3. ある程度は出来る
    4. ほぼできる

  で回答

分析・結果

- 多相ラッシュ分析（Linacre, 2017）
  - 受験者能力、タスク難易度、評価者の厳しさを相として分析
  - スピーキングデータ、アンケートデータを分析
- スコアの全体の分散の58.63%が説明できた。
  - 一次元性を満たした
  - 1つの能力を測っている

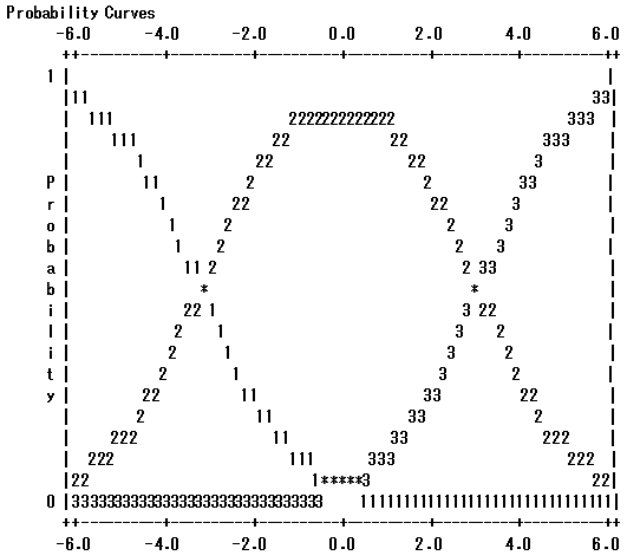
```
+-------------------------------------------------+
|Measr|+Ss       |-Task              |-Rater|Scale|
|-----+----------+------------------+------+-----|
|   4 + *****    +                  +      + (3) |
|     | ***      | A2.1.2           |      |     |
|   3 + *****    +                  +      + --- |
|     | **       | A2.1.1           |      |     |
|   2 + *****    +                  +      +     |
|     | ****     |                  |      |     |
|   1 + *****    + A1.1.1   A1.3.1  +      +     |
|     |          |                  |      |     |
*   0 * ****     * A1.2.2   A1.3.2  * 1  2 *  2  *
|     | *******  | A1.1.2   A1.2.1  |      |     |
|  -1 + ********* +                  +      +     |
|     | ****     |                  |      |     |
|  -2 + **       +                  +      +     |
|     | ****     |                  |      |     |
|  -3 + **       + PreA1.2          +      + --- |
|     | *        | PreA1.1          |      |     |
|  -4 + ****     +                  +      + (1) |
|-----+----------+------------------+------+-----|
|Measr| * = 1    |-Task              |-Rater|Scale|
+-------------------------------------------------+
```

	平均	標準偏差	最小値	最大値	分離(strata)	信頼性
受験者能力	0.25	2.66	-5.73	8.91 (満点)	4.18	.89
タスク難易度	0.00	2.10	-3.69	3.51	9.60	.98
採点者	0.00	0.20	-0.20	0.20	1.75	.53

- 評価者間一致度75.5%（ラッシュ分析での予想70.8%)
- 適合の基準:インフィット平均平方値 = 0.50～1.50 (データがラッシュモデルの予想に適合した回答で問題ないと判断)
- タスク:全て適合（0.66～1.24）
- 受験者: 53.03% が適合
  - 0.5未満は25.76%、1.5より大は18.18%（2.0より大は7.58%）
  - 2.0以上回答を分析。特にテストとしての改善点は見つからず
- 評価者:全て適合（0.99～1.01）

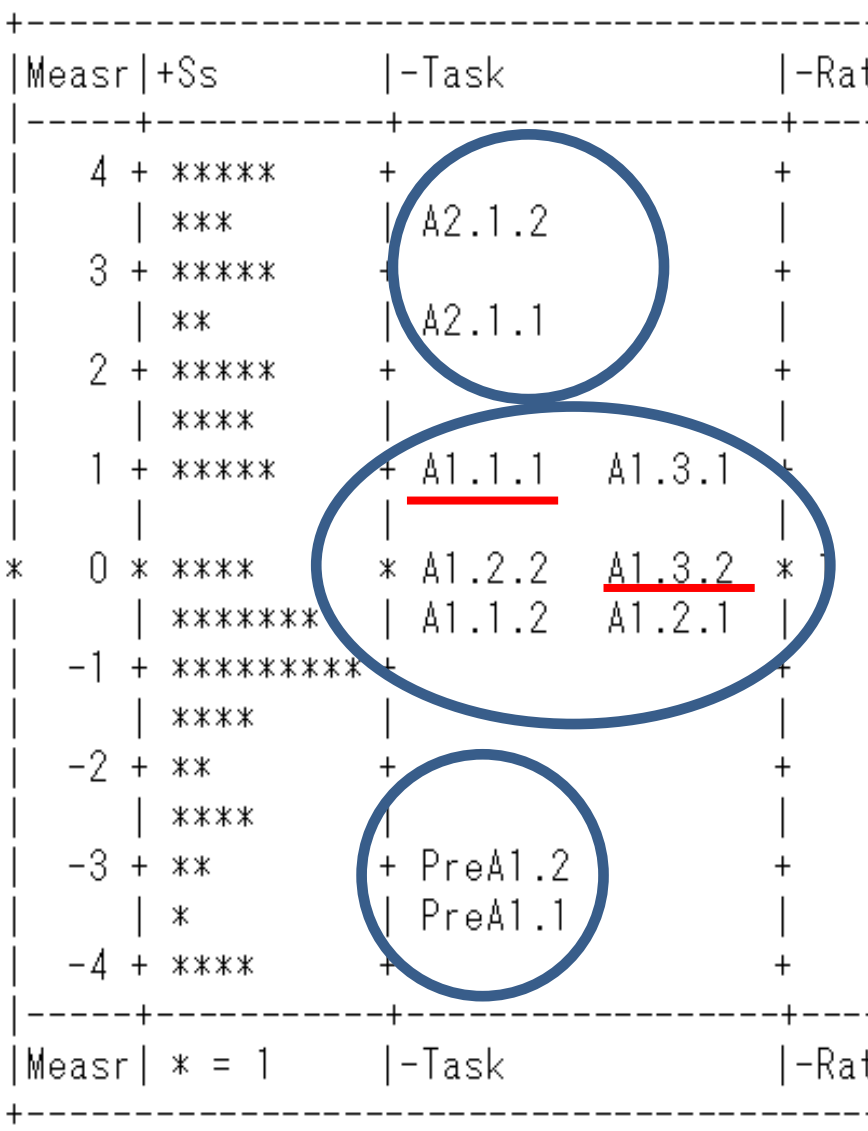
評価基準の適切さ (Bond & Fox, 2015)

- 4観点中3観点を満たしていた
- 要検討事項「敷居 (またはステップ) の間の差 (= 距離) について、隣のレベル同士の距離が1.4ロジット以上、5.0ロジット未満である」
- 結果 : 評定2と3の距離が6.32
- 2と3の間の距離が広く、もう一つ別の基準を入れることも可能
- もしくは、2と3の基準を明確化して、3をもう少し易しくする可能性がある。本調査においての要検討事項。



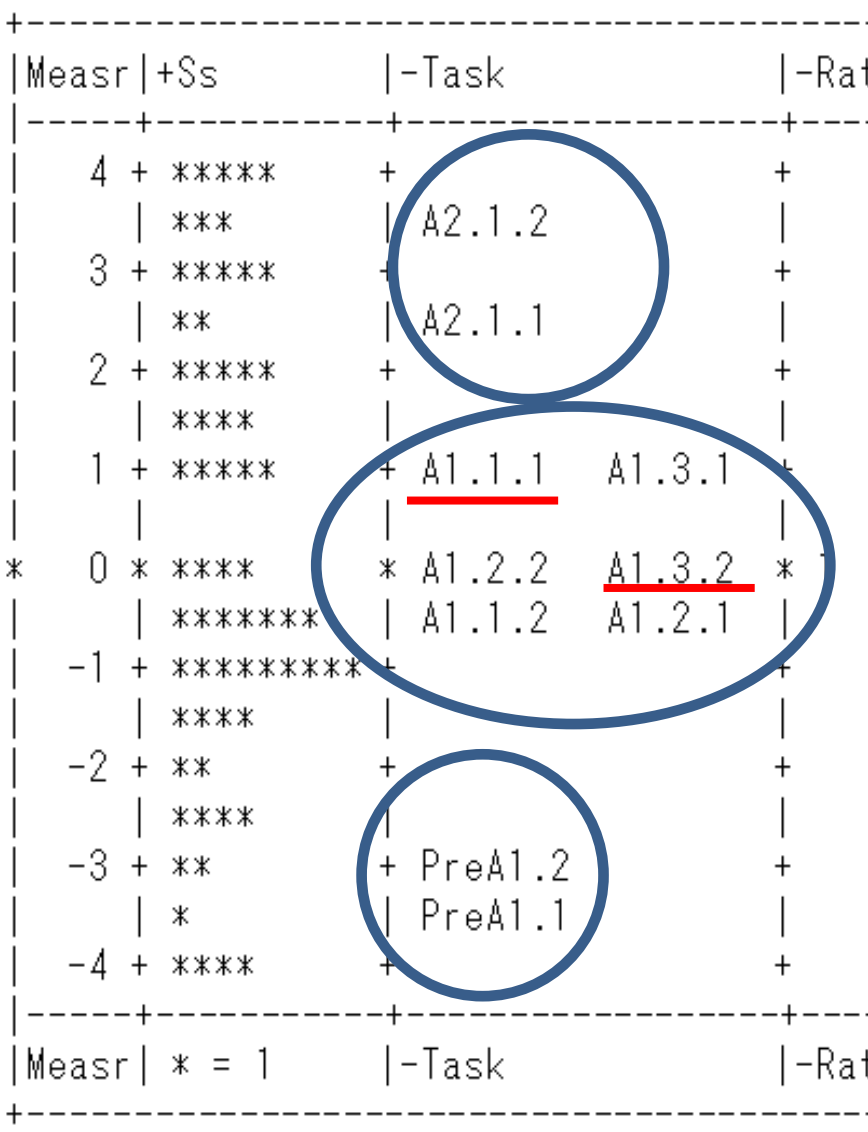
タスクの難易度の順序は意図通りか1/2

- A1.3.2、A1.1.1で予測順と異なる
- 意図した難易度と実際のタスク難易度の関係
  ρ = .86
  - 意図した難易度が高いタスクでは、実際のタスク難易度も高い、という強い傾向
  - 全体的には意図と結果は一致
- 誤差を考慮すると、難易度グループは3つ



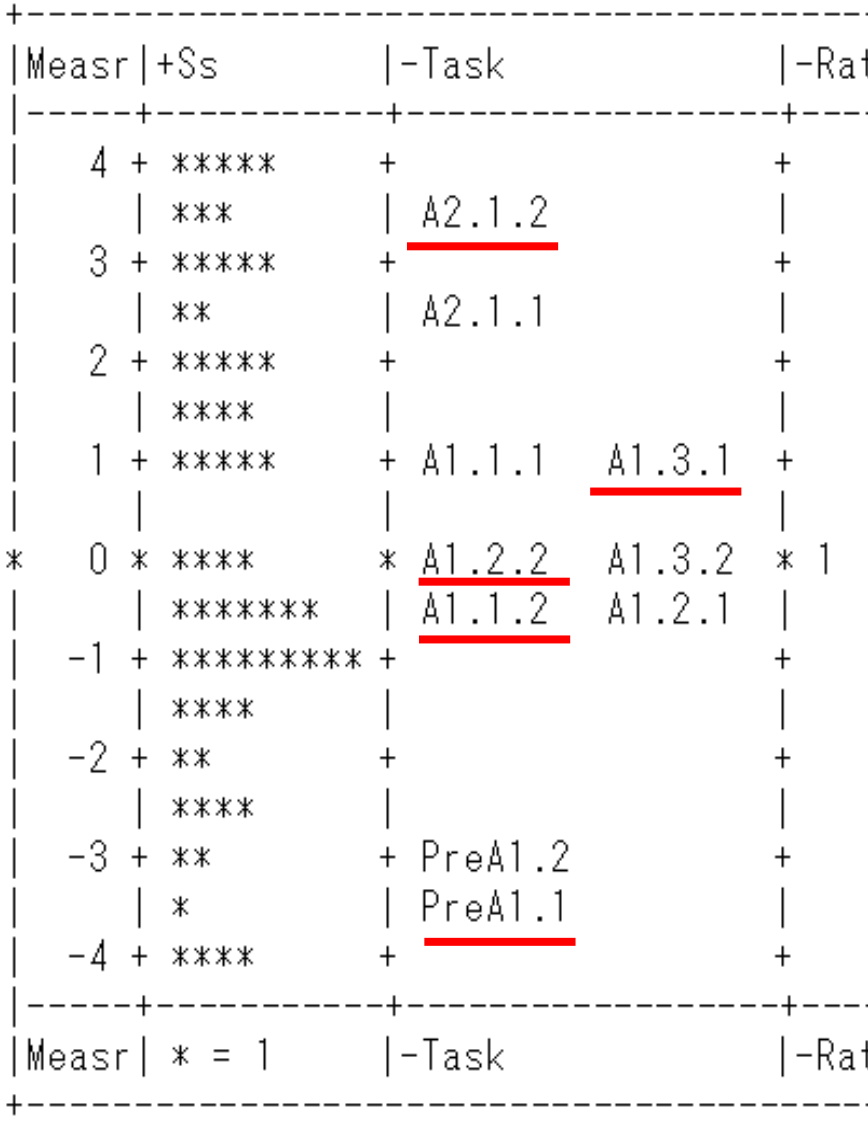
タスクの難易度の順序は意図通りか 2/2

- A1.3.2、A1.1.1で予測順と異なる
- A1.1の1個目：information gapのあるカードを使った、イベントでの情報をやりとり
  - A1.1の記述子とタスクの対応の確認
  - 評価基準の改善点：A1.1で求めるText typeを明示。
- A1.3の2個目：誘われたが断り、その後受けるタスク
  - 評価基準の改善点：A1.3で求めるText typeを明示



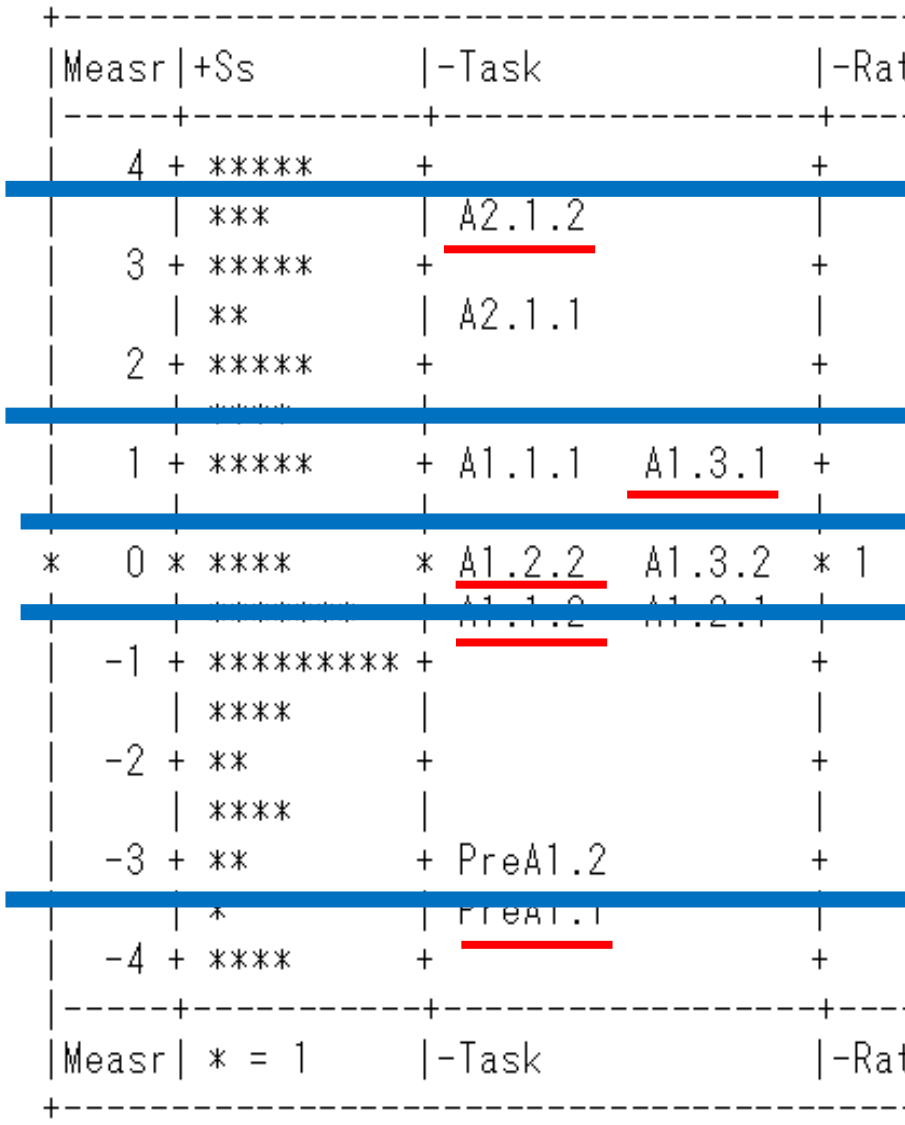
各CEFR-Jレベルに受験者はどの程度位置づけられるか

- 代表性のあるタスクを各レベル1個選んだ
- Pre-A1.1、A1.1.2、A1.2.2、A1.3.1、A2.1.2 のタスクをかなりの確率で到達できている　＝　このレベルをパス
- 60%で到達（logit = +0.4）、70%で到達（logit = +0.8）、80%で到達（logit = +1.4）として、人数を数えた



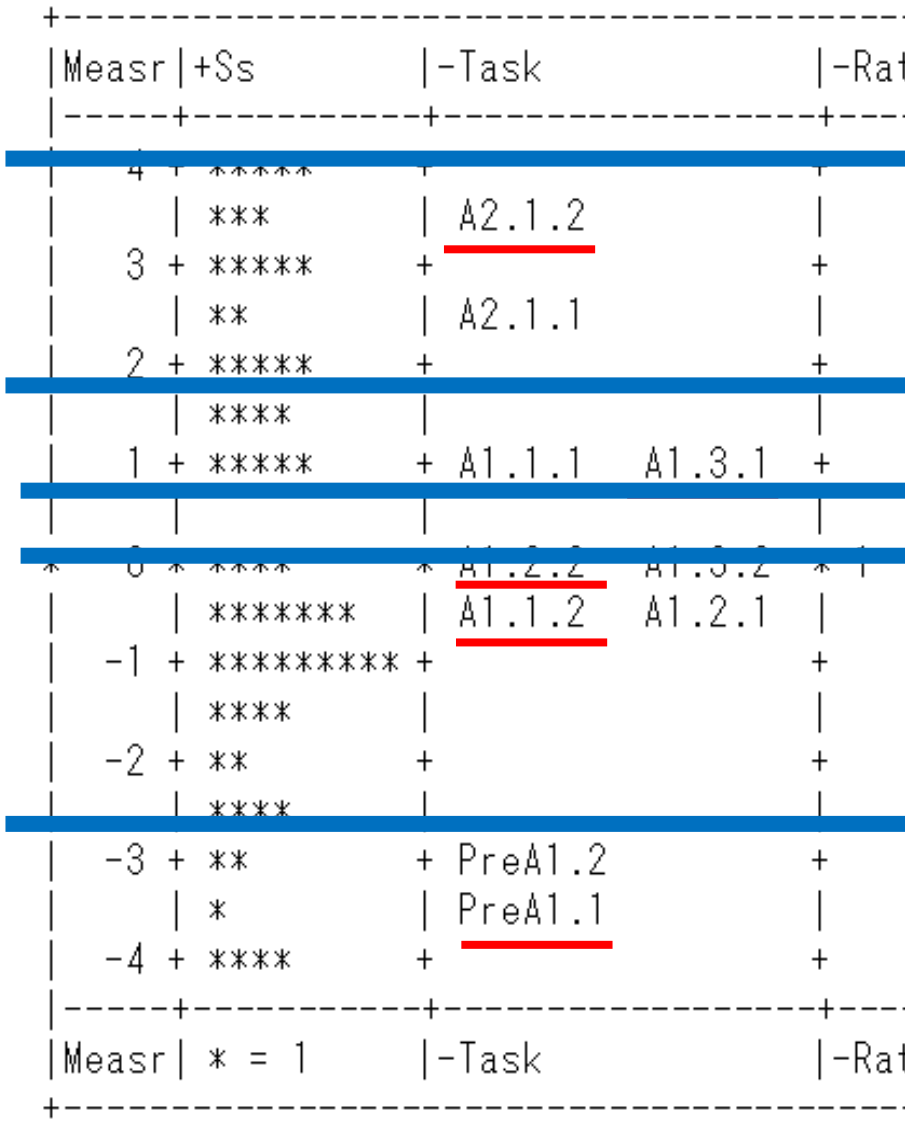
60%で到達のときの基準点

- logit = +0.4　A2.1以上
- A1.3
- A1.2
- A1.1
- Pre-A1
- Pre-A1未満



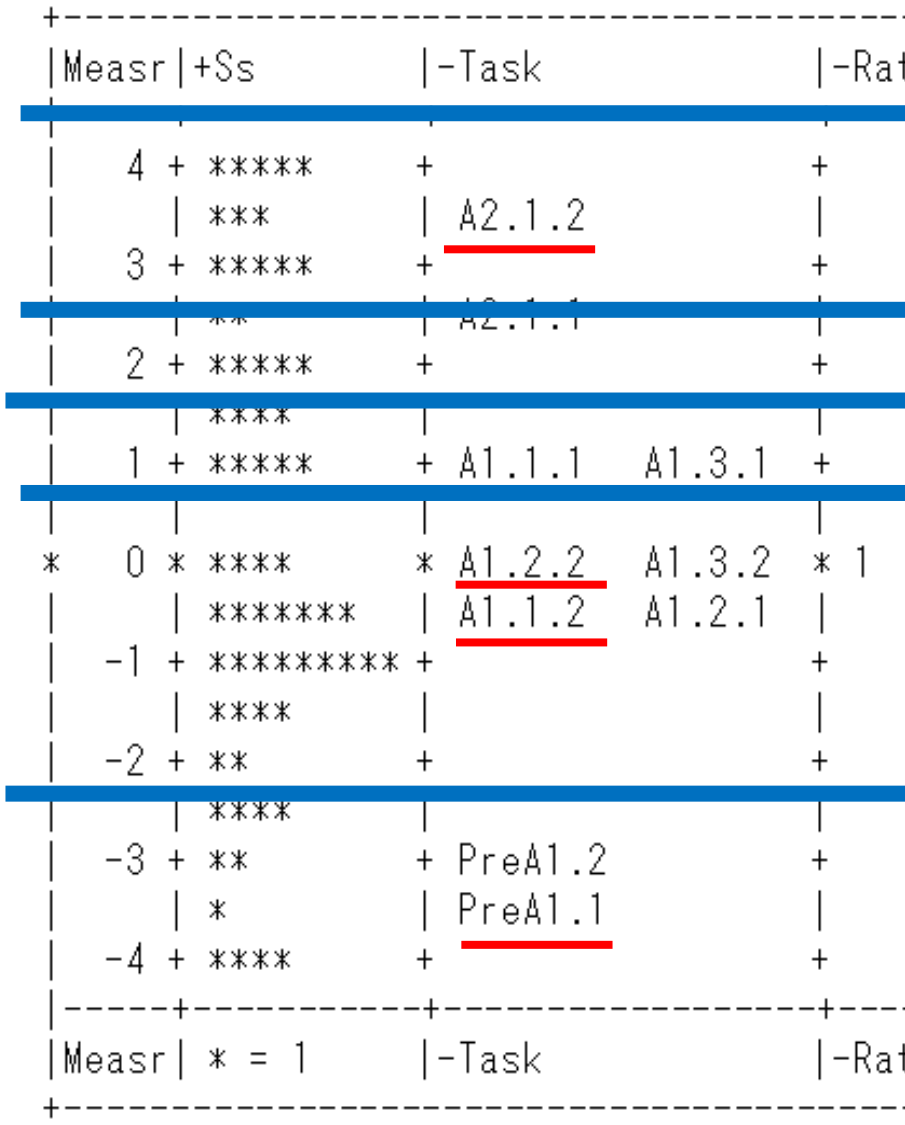
70%で到達のときの基準点

- logit = +0.8　A2.1以上
- A1.3
- A1.2
- A1.1
- Pre-A1
- Pre-A1未満

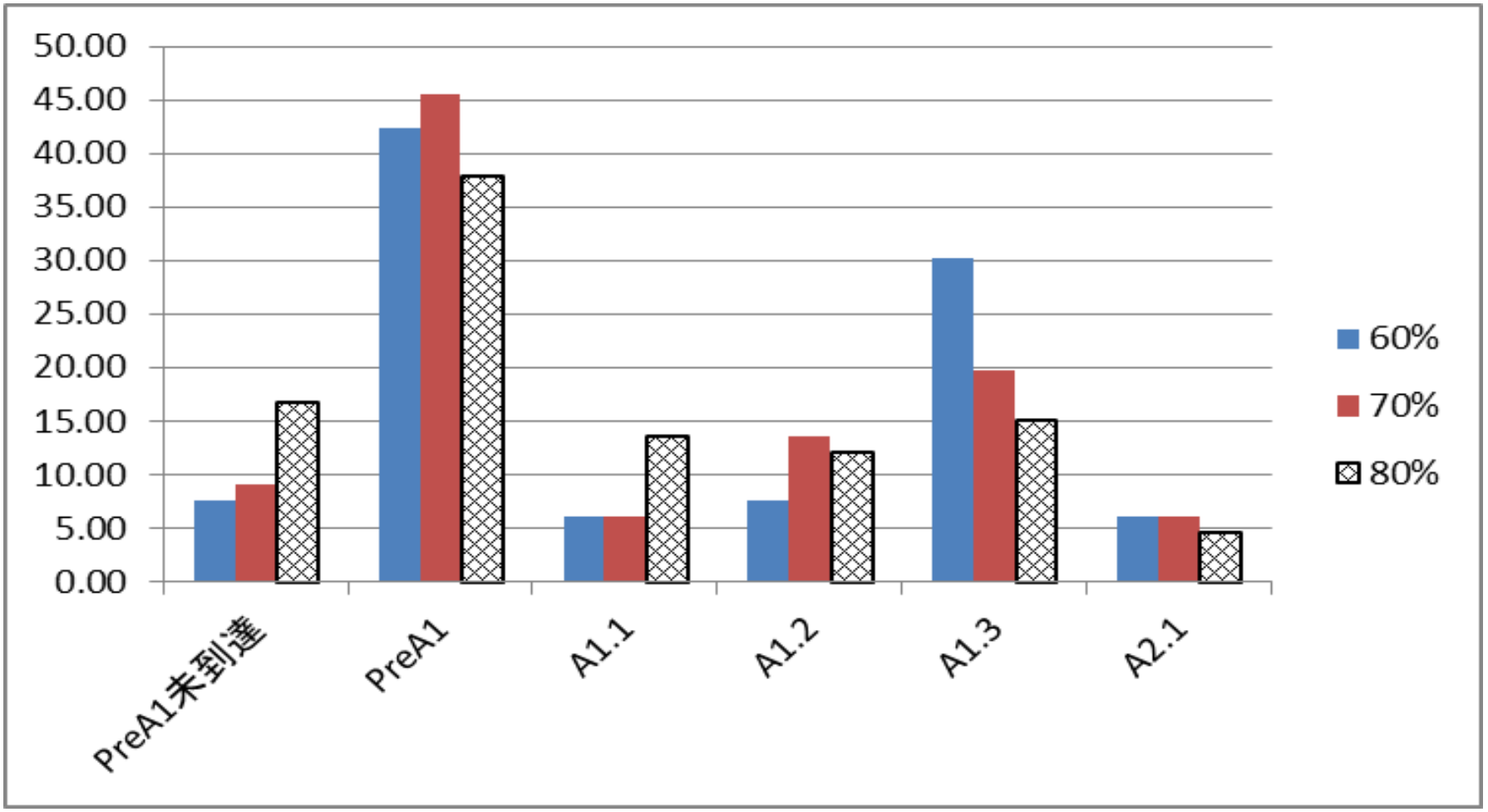


80%で到達のときの基準点

- logit = +1.4　A2.1以上
- A1.3
- A1.2
- A1.1
- Pre-A1
- Pre-A1未満

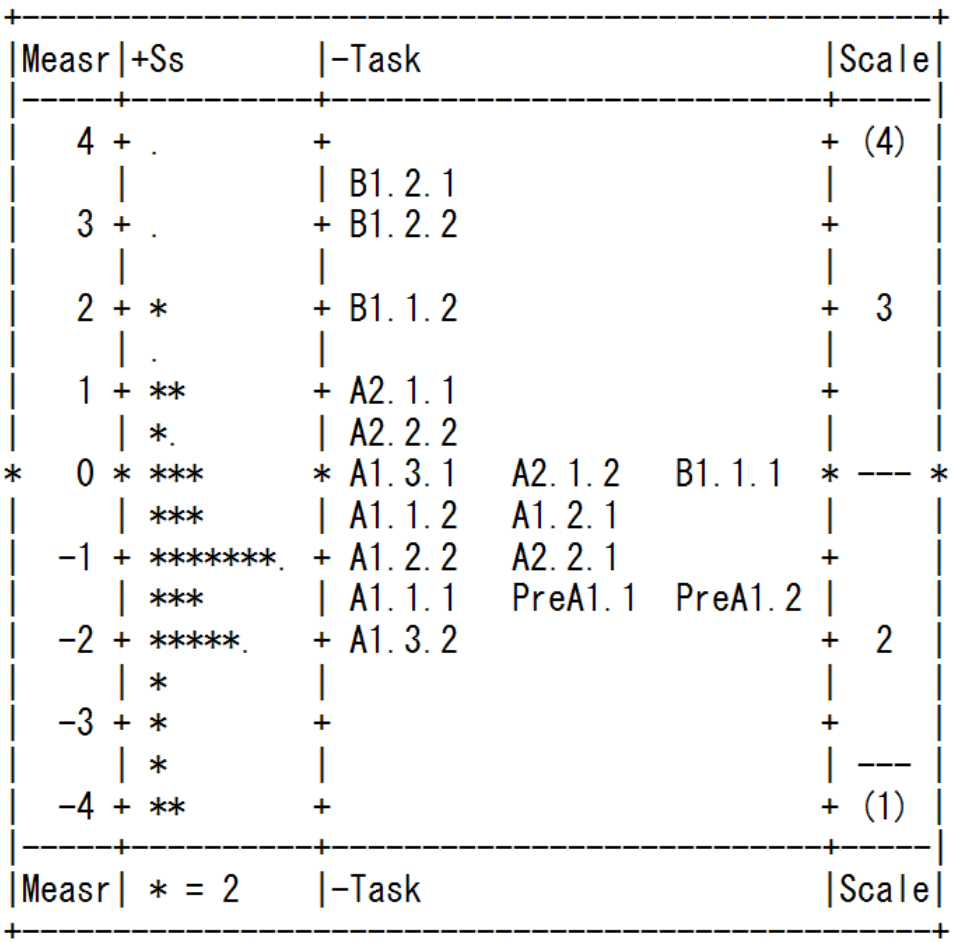


各CEFR-Jレベルに位置づけられた受験者の割合



アンケート分析

- 項目すべて適合
- CEFR-Jで意図した順位と実際の順位の関係：ρ = .79
- やりとり力のテスト結果と自己評価の関係：*r* = .46（中程度）
  - 測りたい力が測れている傾向。テストで測るやりとり力と自己評価で測るやりとり力は、同じ能力だが方法が異なる。望ましい結果

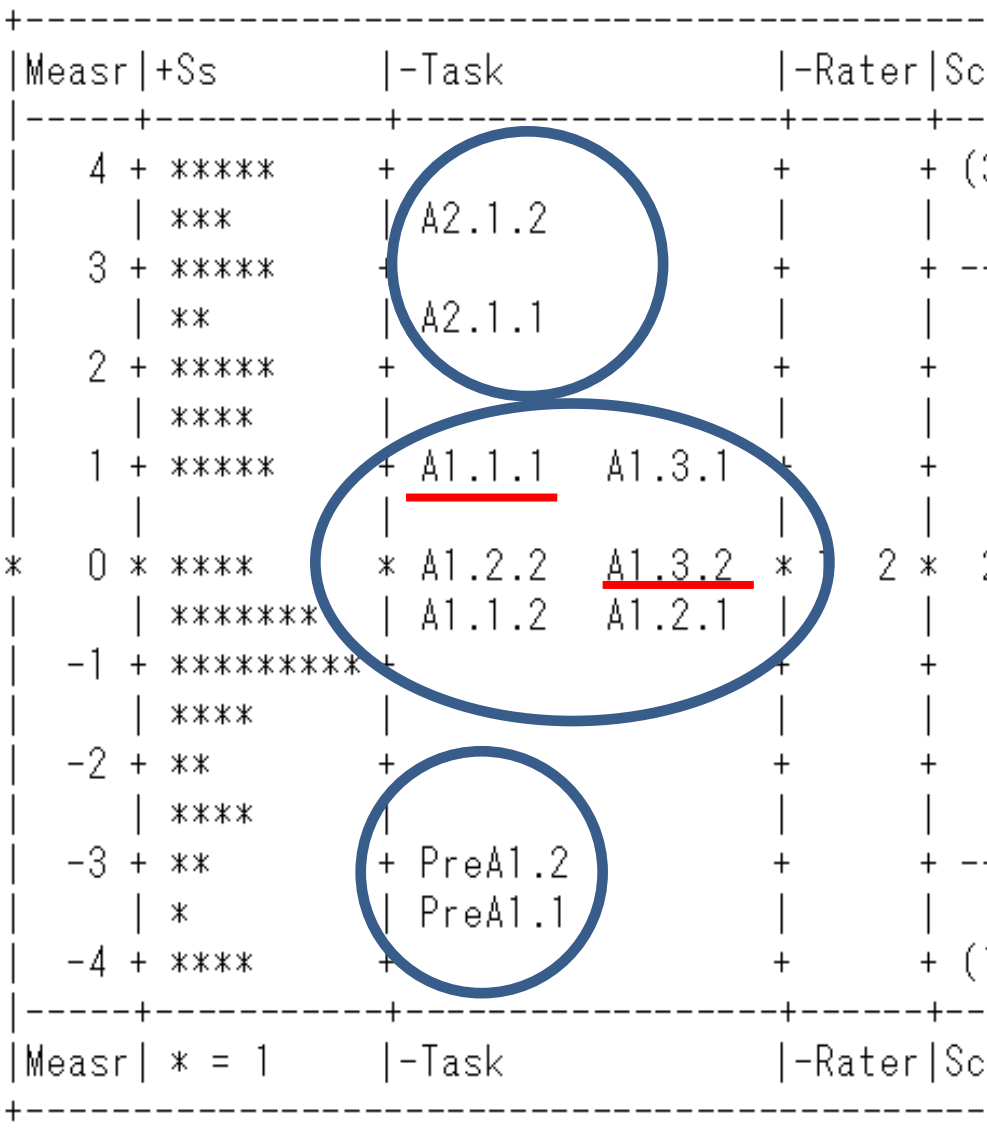


今後の方向性

- 受験者人数を増やす
- タスクの改善
- 評価基準の改善
- 回答をすべて2名以上で採点し、採点の安定性（信頼性）を確認
- その他古典的テスト理論での分析と合わせて、多相ラッシュ分析を行う

まとめ

- PreA1, A1.1, A1.2, A1.3, A2.1の5レベル10タスク実施
- A1.3.2、A1.1.1で予測順と異なるが全体的な傾向は一致
- やりとり力のテスト結果と自己評価の関係：中程度
  - 測りたいやりとり力が測れている傾向



引用文献

- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.
- Linacre, J. M. (2017). Facets: Many-Facet Rasch-measurement (Version 3.80.0) [Computer software]. Chicago: MESA Press.