Measuring, Modeling, and Shaping Skill Development

Andrew Caplin: HCEO Conference on Measuring and Assessing Skills

Chicago, October 2 2015

Introduction

- Will pose five basic (abstract) questions
- Question 1: How well does standard multiple choice test with standard grading measure skill?
  - 1A: How is standard test answered?
  - 1B: What therefore can be inferred from scores?
- Question 2: Data engineer's question: how might enriched measurement and grading improve skill measurement?
  - 2A: Elicit information about confidence in answer and use in grading algorithm
  - 2B: Elicit information about (or restrict) allocation of time and use in grading algorithm
- Question 3: How would changes in measurement and scoring impact learning?

Introduction

- Brief answers to Q1-Q3:
- Question 1: How well does standard multiple choice test with standard grading measure skill?
- Use simple e.g.s to illustrate reasons to worry
  - In simplest reasonable model, mapping from beliefs about answers to answer depends on scoring rule and utility function
  - In simplest reasonable model, optimal allocation of time problem essentially insoluble
  - In richer model, role for psychological variables (e.g. anxiety)

Introduction

- Question 2: How might enriched measurement and grading improve skill measurement?
- Use simple e.g.s to illustrate reasons for optimism
    - In simplest reasonable model allowing elimination and eliciting beliefs revealing
    - In simplest reasonable model much learned from allocation of time revealing
    - Measuring both even richer
    - Improves adaptive testing in vertical learning environments

Introduction

- Question 3: How would changes in measurement and scoring impact learning?
  - In given exam, test taker (TT) with fixed actual skill (cognitive capacity) must map from prior learning to distribution of possible scores and corresponding utilities
  - Extremely complex since scores based on posterior beliefs which depend on time allocation
  - Best possible posterior depends on grading scheme and external value
  - TT has beliefs about distribution of possible tests
  - This allows computation of EU of any given level of skill

Introduction

- Balance utility of capacity against costs
  - TT has utility costs (time, effort, and angst) of skill development
  - Based on some view of the personal production function for cog. capacity chooses optimal level of such development!
  - Not at all easy to specify
  - Hints from theory of rational inattention (Sims [1998, 2003], Woodford [2012], Matejka and McKay [2015], Caplin and Dean [2015]).

Introduction

- Question 4: What research methods would liberate further understanding?
  - I propose a class of laboratory experiments before field tests
  - Simple idea is to fix skill by fiat and explore how well measured in different protocols.
  - Can enforce different time divisions to get sense of feasible set of posteriors
  - Can add ex ante purchase to get to the investment phase
- Note no attempt to introduce theory of optimal design at this point
  - A bridge too far

Q1A: Knowledge and Score

- 1A: How is standard test answered?
- First part is how does examinee knowledge at point of completion impact answers?
- **Standard MC test** M has three parameters:
  - T time (minutes) available to answer all questions
  - N no. of distinct questions drawn from $q(n) \in Q$ background question set;
  - $K \geq 2$ real answer options per question

Q1A: Knowledge and Score

- Action set for each question is Y:

$$Y = \{1, , , K, \emptyset\};$$

  with $\emptyset$ denoting no answer.
- Actual answer (in words) associated with option k for question n is $a(k, n)$ from universal answer set A
- Unique correct action for each question $y^*(n) \in \{1, , , K\}$
- Typically uniform probability independent across questions in the design that each is correct.

Q1A: Knowledge and Score

- A **standard answer** is an element of $\bar{y} = (y(n))_{n=1}^{N} \in Y^N$.
- **A standard scoring rule** is a piece-wise linear function $\sigma : Y^N \longrightarrow [0, N]$ depending only on the number of correct and incorrect answers

$$
\begin{aligned}
C(\bar{y}) &= \sum_{n=1}^{N} 1_{\{y(n)=y^*(n)\}}; \\
I(\bar{y}) &= N - C(\bar{y}) - \sum_{n=1}^{N} 1_{\{y(n)=\emptyset\}}; \\
\sigma(\bar{y}) &= \max\{C(\bar{y}) - \rho I(\bar{y}), 0\};
\end{aligned}
$$

with $\rho \geq 0$ the error penalty.

Q1A: Knowledge and Score

- Test given to individuals $i \in I$; with $\bar{y}^i \in Y^N$ the answer of i and $\sigma(\bar{y}^i)$ the corresponding score.
- What examiner learns about $i \in I$ depends on what determines these answers
- Here we enter realm of theory

Q1A: Knowledge and Score

- Simplest reasonable model a Bayesian maximizing expected utility of the final score,
$$U : [0, N] \longrightarrow \mathbb{R}.$$
- To formalize define posterior beliefs at point of choosing all answers that $\bar{y} \in [Y/\emptyset]^N$ is correct vector of answers: must sum to 1.
- Correlations can be induced by common aspects of answer algorithm.
- Optimal answer problem non-trivial
- This treats it as all answered at once at end: equivalent if can go back and change in light of noted correlations
  - Else even more complex
  - Standard batch vs. sequential issue in search theory

Q1A: Knowledge and Score

- Simplest is independent case (sequential and batch answer strategies the same)
- Define $\gamma^i(k, n)$ as i's posterior at point of answer that $1 \leq k \leq K$ is correct answer to question $1 \leq n \leq N$.
- In independent case, if answer, surely pick some most likely element $\hat{k}(n)$ (for simplicity unique)

$$y^i(n) \in \arg\max_{1 \leq k \leq K} \gamma^i(k, n) \cup \emptyset.$$

Q1A: Knowledge and Score

- When best to not answer?
- Simple(st?) theory would be a threshold rule based on posterior beliefs over the correct answers to each question.
- Simplest satisficing rule is to set penalty dependent threshold probability $\bar{\gamma}(\rho)$ and answer

$$\begin{aligned}
\max_{1 \le k \le K} \gamma^i(k, n) &\ge \bar{\gamma}(\rho) \Longrightarrow y^i(n) \in \arg \max_{1 \le k \le K} \gamma^i(k, n); \\
\max_{1 \le k \le K} \gamma^i(k, n) &< \bar{\gamma}(\rho) \Longrightarrow y^i(n) = \emptyset.
\end{aligned}$$

- Defines complete mapping from posteriors to possible answers.

Q1A: Knowledge and Score

- Relies on linear EU over score
  - Inconsistent with floor of 0
- A risk averter may get all "most likely correct" to probability $p > \frac{1}{K}$ correct but find it better to not answer some if this lowers the probability of catastrophic outcome
  - e.g. three questions penalty $\rho > 0$ and need to get at least 2 to avoid catastrophe
  - If answer 2 get 2 probability p^2: answering all 3 dominated since need to get all three right to avoid catastrophe, probability p^3.
- In independent case general optimal strategy based on posterior is to look at EU if answer first m most likely and then do not answer rest.
- Call this $V(m)$ and then maximize over m.

Q1A: Knowledge and Score

- With correlated answers get choice between plunging and diversification
- Two answer algorithms each 0.5 correct determine answer to 2 questions
  - Get 2 questions, no (small) error penalty and concave EU: alternate answers
  - If need both correct for EU reasons then instead plunge
- Qualitatively: may need to change prior answer to optimize given evolving information about correlations

Q1A: Knowledge and Score

- Above gives no role to time allocation and time constraint
  - Drift-diffusion model (Ratclifff[1978]) shows that more time generally raises probability correct.
- Hence score depends on time allocation strategy
  - Easy first beats linear order: different form of intelligence to know
  - Caplin and Martin [2015] experiment shows bi-modal time to decide:
  - Quick decision guess or not:
    - If guess look like only trivial information taken in
    - If not, deliberate and to better

Q1A: Knowledge and Score

- What best stopping time for identifying hard question and what to do with that?
- Depends on what happens next: essentially impossible dynamic programming problem!
- Psychological characteristics also enter:
  - How early problem impacts later performance may depend on neuroticism

Q1B: Score and Skill

- What then to infer from scores?
- If RE and beliefs correct on average ($p = 0.9$ is 90% correct) then if all answered with same confidence, score a good estimator as number of questions increases
- Can define more skilled type as one who is more certain about the answers to all questions
- Induces a mapping, albeit stochastic, from skill to score distribution
- Underlies simple theory that higher score likely reflects higher skill.

Q1B: Score and Skill

- But in richer and more realistic theory conflates many factors:
  - With non-linear EU may answer more if less confident and produce higher expected score.
  - Different utility functions possible so score reflects preferences and skill:
  - Character differences e.g. anxiety
  - Illusory beliefs e.g. overconfidence ($p = 0.9$ is 60% correct)
- Might find an individual who dominates another in sense of clarity per unit time yet scores lower
  - Different order of answers
  - Different cutoff strategy (too much time on a hard question)

Q2A: Posteriors and Elimination

- Simple schemes can recover more details of posterior
  - If allow at least occasionally multiple options and/or elimination
- In principle may measure actual posteriors of most likely
  - BDM scheme for replacing 1 based on belief draw: use question if draw lower than stated belief and else use stated belief and urn!
  - Enables test of RE: may reveal possibly dangerous illusion of certainty!
  - Interesting question of whether or not to allow no score: maybe want this but also most likely if forced again with BDM
  - To get out information on correlations in beliefs requires conditional probabilities!
- Measuring beliefs may allow separation of "Eureka" from continuous accretion questions

Q2B: Time

- With time allocation can do better skill identification
- Can use an interface that enforces order and removes differences in the strategy.
  - Makes it a more direct reflection of task skill
- If want to know about skill in selection algorithm, design a separate test!

Q2B: Adaptive Testing

- Exam design very different vertical in difficulty vs. horizontal (all equally difficult)
- Superior measurement improves adaptive testing in vertical cases.
  - Not just errors but remaining time
  - Provides possibility for interactive hints as time extends

Q3: Optimal Development and Deployment of Skill

- First fix exam protocol and grading scheme
- Fixed actual skill (cognitive capacity: think Shannon capacity as example) determined by pre-exam effort (see below)
- Also an EU function over scores based on value in future options/career
- In given multiple choice test $M \in \mathcal{M}$, reasonable that test taker (TT) has unifom prior over correct answers
- Utility function induces mapping from vector of posteriors to answers to scores

Q3: Optimal Development and Deployment of Skill

- Designing an information system in the sense of Blackwell
  - Essentially a mapping from the uniform prior to a distribution over possible posteriors.
  - Can formulate as a classical optimization problem in language of RI
- The true answers are hard to assess: the goal of the TT is to choose a clarifying information structure using fixed skill
  - Depending on time allocation will end up with different profile of posteriors and hence optimal answers and scores
  - TT might identify optimal exploration and answer strategy in non-anticipatory manner
- RI appropriate to focus on internal cognitive constraints on information processing rather than external costs of information access.

Q3: Optimal Development and Deployment of Skill

- The learner's job ex ante is to invest in earning a valuable score subject to the individual costs of building this skill
- From an ex ante view the actual learning during pre-exam period motivated not by given exam but by beliefs over the exam
- From ex ante viewpoint must judge how skill level impacts score on all possible tests
  - Think of investment in capacity in relation to the larger space of all possible questions and their answers.
  - Requires beliefs about possible exams as set by the teacher (will not look for consistency now!)
  - This allows computation of EU of any given level of skill

Q3: Optimal Development and Deployment of Skill

- It is envisaged that capacity is subjectively costly to produce.
- In basic RI theory, the DM faced with maximizes expected utility net of (separable) capacity costs.
  - Different RI models involve differentially specifying the notion of capacity and the cost function for building it
  - Of particular importance is the Shannon cost function which specifies costs as linear Shannon capacity
- To a first approximation, goal of exam is to encourage the building of the capacity
  - Examiner's optimization a bridge too far

Q4: Experimental Elicitation of Skill

- Question 4: What research methods would liberate further understanding?
    - Fix skill: make questions involve various operations carried out by a machine.
    - Make one machine faster in all operations by a fixed proportion
    - Have them complete a large set of different types of test
    - See how well you can recover fixed skill
    - To induce emotions make difficult tasks hard to identify
    - Do a personality inventory etc. to see how other factors enter.