Difficulty, item, and answer modeling: An intersection of AI and assessment

Isaac I. Bejar and Michael Flor

ETS

17th Annual Maryland Conference

University of Maryland, College park

November 2-3, 2017

11/13/2017

Difficulty, item, and answer modeling: An intersection of AI and assessment

Isaac I. Bejar and Michael Flor

ETS

17th Annual Maryland Conference

University of Maryland, College park

November 2-3, 2017

Goal and overview:

- Explore the benefits of recent AI research, broadly conceived, for psychometrics, specifically the design of tests
- Overview
  - Background on AI
  - Item generation and difficulty modeling
  - Answer modeling
  - Applications to GRE®
  - Conclusions

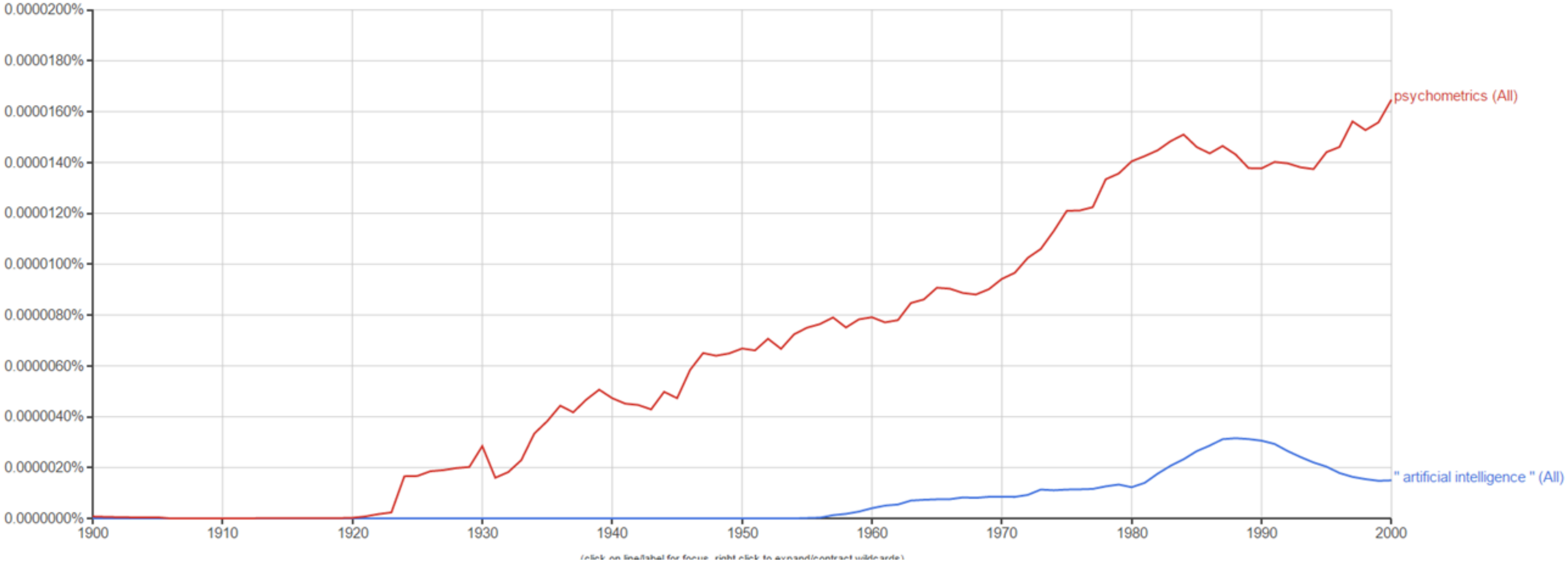
Background on AI

Artificial intelligence

THE QUEST FOR ARTIFICIAL INTELLIGENCE
A HISTORY OF IDEAS AND ACHIEVEMENTS
Nils J. Nilsson
Stanford University

- The website for the AAAI state as its goal is the

  “advancing the scientific understanding of the mechanisms underlying thought and intelligent behavior *and their embodiment in machines*.”



Psychology in the 1950's

While experimental psychologists were rethinking the definition of psychology, other important developments were occurring elsewhere. Norbert Wiener's cybernetics was gaining popularity, Marvin Minsky and John McCarthy were inventing artificial intelligence, and Alan Newell and Herb Simon were using computers to simulate cognitive processes. Finally, Chomsky was single-handedly redefining linguistics.(p. 142)

Miller, G. (2003). The cognitive revolution: a historical perspective. Trends in Cognitive Sciences, 7, (3), 141-144.

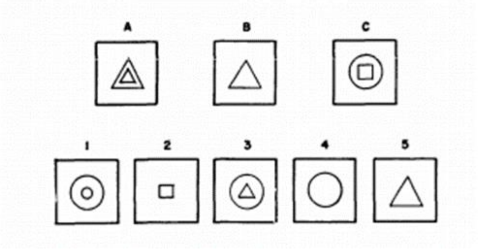
Not to be left behind…..

The work of Lord (1952) and (intellectually) closely related work by mathematical sociologist Paul Lazarsfeld (1950) clarified the nature of IRT models and emphasized the idea that they were "latent variable" models that explained the observed covariation among item responses….(p. 12)

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Ed.), Handbook of Statistics: Psychometrics (Vol. 26) (pp. 1-27). New York: Elsevier.

Approaches to evaluating machine intelligence

- The Turing test
  - A behavioristic criterion for determining machine intelligence
- Performance on cognitive tasks: answer modeling
  - Evans, T. G. (1968). Program for the solution of a class of geometry-analogy intelligent-test questions. In M. Minsky (Ed.), Semantic information processing MA: MIT Press. (Based on 1963 thesis.)



Item taken from the 1942 edition of the Psychological Test for College Freshmen of the American Council on Education

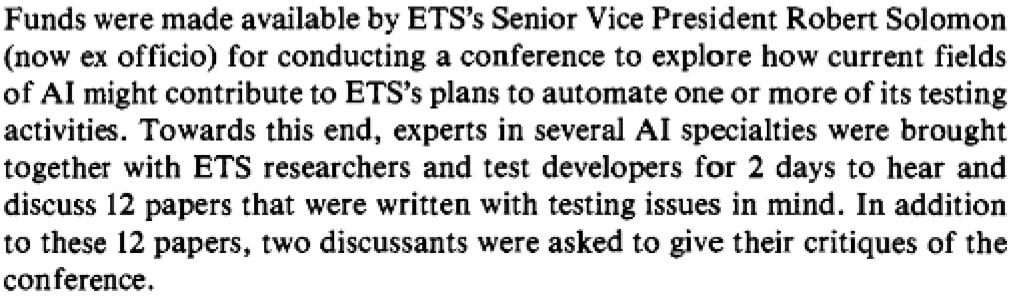
Measuring the Power of Learning.®

Psychometric issues

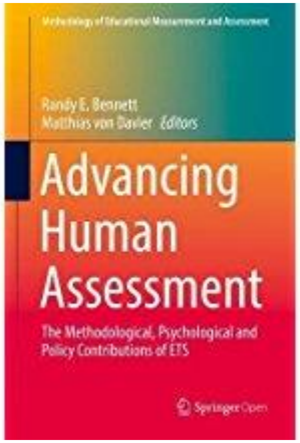
- Desired inference: The machine performs well on a set of tasks that require intelligence, therefore the machine is intelligent
  - Which tasks?
    - Are they a sample from a well-defined universe?
    - What kind of sample?
    - Would performance generalize to different samples?
  - If not all items are answered, what is the nature of the answered and unanswered items?

Early encounter with AI at ETS

Freedle, R. O. (1990). *Artificial intelligence and the future of testing.* Hillsdale, NJ: Lawrence Erlbaum Associates.



Funds were made available by ETS's Senior Vice President Robert Solomon (now ex officio) for conducting a conference to explore how current fields of AI might contribute to ETS's plans to automate one or more of its testing activities. Towards this end, experts in several AI specialties were brought together with ETS researchers and test developers for 2 days to hear and discuss 12 papers that were written with testing issues in mind. In addition to these 12 papers, two discussants were asked to give their critiques of the conference.


Methodology of Educational Measurement and Assessment
Randy E. Bennett
Matthias von Davier Editors
Advancing Human Assessment
The Methodological, Psychological and Policy Contributions of ETS
Springer Open

Free e-book!

Increasing recognition of psychometric perspective

- Bringsjord, S., & Schimanski, B. (2003). What is artificial intelligence? Psychometric AI as an answer. Proceedings of the 18th international joint conference on Artificial intelligence, 887-893.
- Hernández-Orallo, J. (2016, August 19). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artificial Intelligence Review.
- Weston, J., Bordes, A., Chopra, S., & Mikolov, T. (2015). Towards AI-complete question answering: a set of prerequisite toy tasks. https://arxiv.org/abs/1502.05698.
- Chmait, N., Dowe, D. L., Li, Y.-F., & Green, D. G. (2017). An Information-Theoretic Predictive Model for the Accuracy of AI Agents Adapted from Psychometrics. In T. Everitt, B. Goertzel, & A. Potapov (Ed.), Artificial General Intelligence: 10th International Conference, AGI 2017, Melbourne, VIC, Australia, August 15-18, 2017, Proceedings (pp. 225-236). Cham: Springer International Publishing.
- Clark, P., & Etzioni, O. (2016). My Computer Is an Honor Student —But How Intelligent Is It? Standardized Tests as a Measure of AI. Artificial Intelligence, (Spring), 5-12.

Conclusions

- Explore the applicability of answer modeling
  - Potential applications in
    - Item generation
    - Difficulty modeling
    - Item analysis, especially distractor analysis

Item generation and difficulty modeling

Item generation: Validity and efficiency

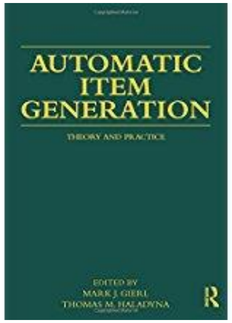
- Construct representation
- Construct preservation

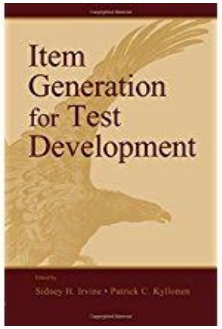
Guttman, L., & Schlesinger, I. M. (1967). Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement, 27,* 569-580.

Hively, W., Patterson, H. L., & Page , S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5,* 275-290.

Bormuth, J. R. (1970). *On the theory of achievement test items.* Chicago, IL: University of Chicago Press.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematics skills. *Cognitive Science, 2,* 155-192.



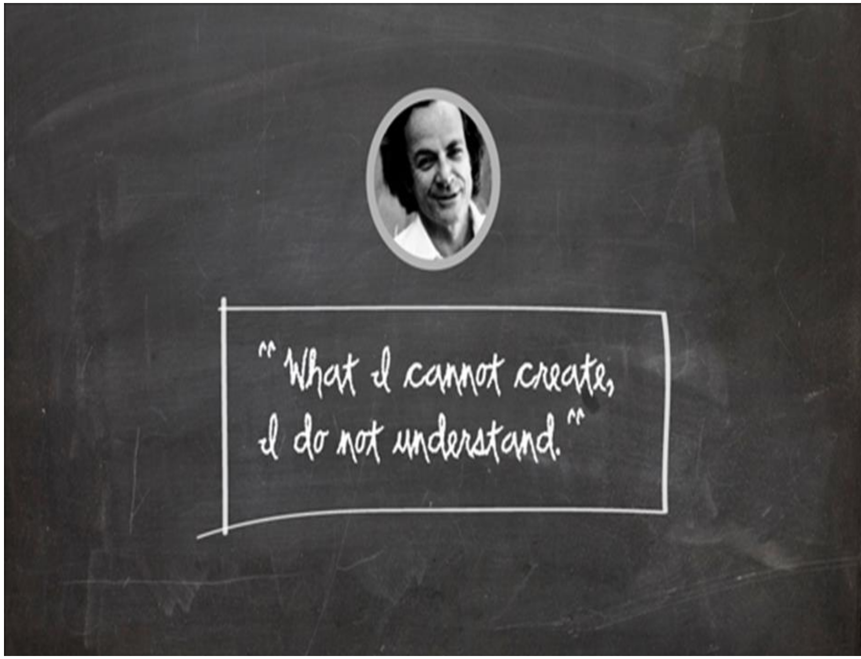

Item Generation for Test Development
Edited by
Sidney H. Irvine • Patrick C. Kyllonen

Susan Embretson

Construct representation

That is, the processes, knowledge stores and strategies that are involved in item solving determine what latent construct(s) are measured by item performance. Implementing the cognitive design system approach involves studying the cognitive components of item solving prior to test development. A set of design principles for items results from developing a plausible cognitive model of the item type. The identified components, along with their associated item stimulus features, are the basis for the item specifications. These design principles predict not only task difficulty but also the specific source of cognitive complexity in each item. (Embretson, 1999, p. 409)

Richard Feynman



While not easy, the rewards of bottom-up approaches are high, as defining the minimal set of components for any process or structure can provide tremendous molecular and mechanistic insights into the cell and how it works.

Way, M. (2017) "What I cannot create, I do not understand". *Journal of Cell Science, 130*, 2941-2942. doi: 10.1242/jcs.209791

Construct preservation

The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it" (Cronbach, 1980, p. 103).

Generativity and Reusability

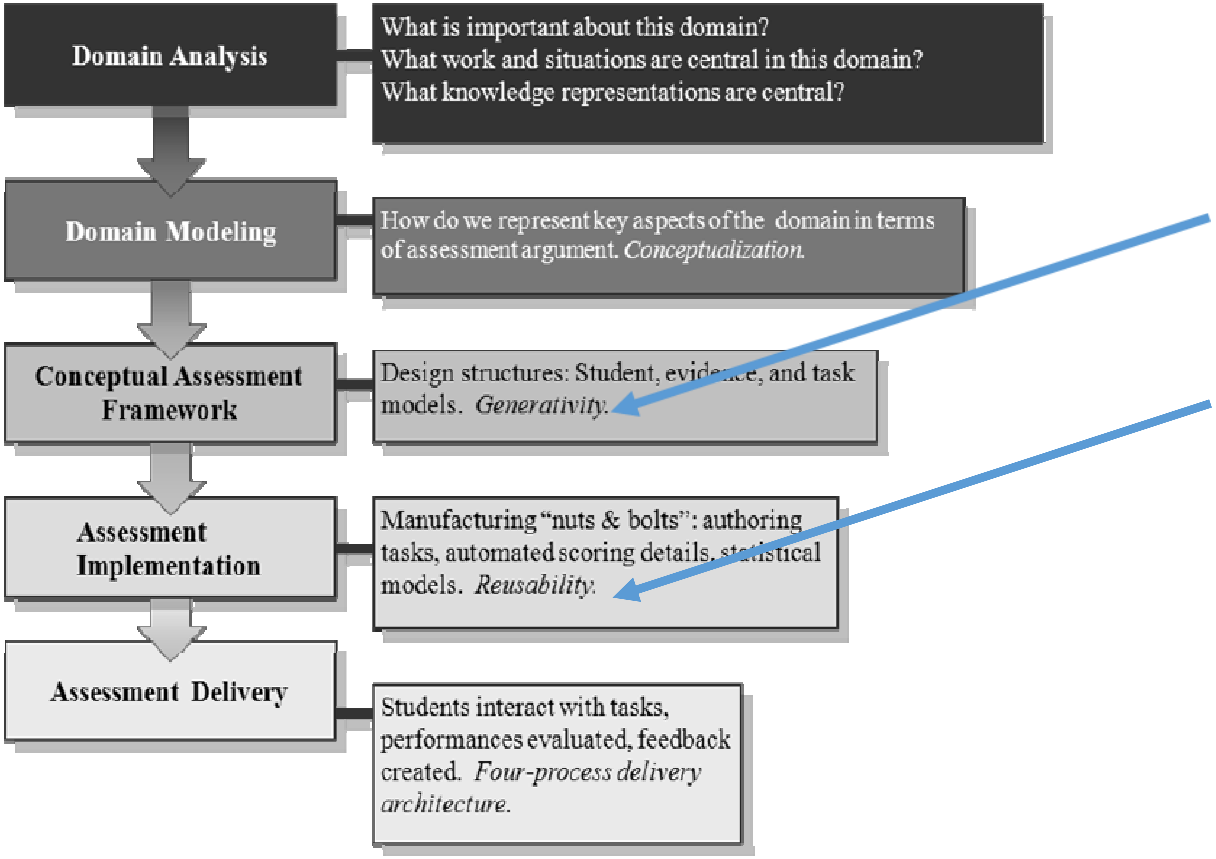
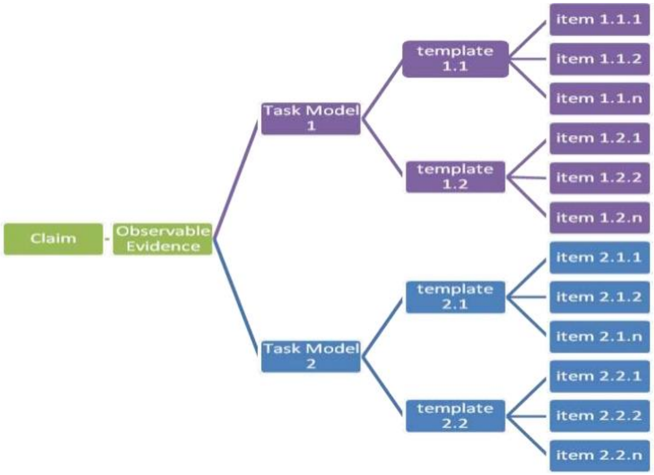

Domain Analysis
What is important about this domain?
What work and situations are central in this domain?
What knowledge representations are central?
Domain Modeling
How do we represent key aspects of the domain in terms of assessment argument. *Conceptualization.*
Conceptual Assessment Framework
Design structures: Student, evidence, and task models. *Generativity.*
Assessment Implementation
Manufacturing "nuts & bolts": authoring tasks, automated scoring details, statistical models. *Reusability.*
Assessment Delivery
Students interact with tasks, performances evaluated, feedback created. *Four-process delivery architecture.*

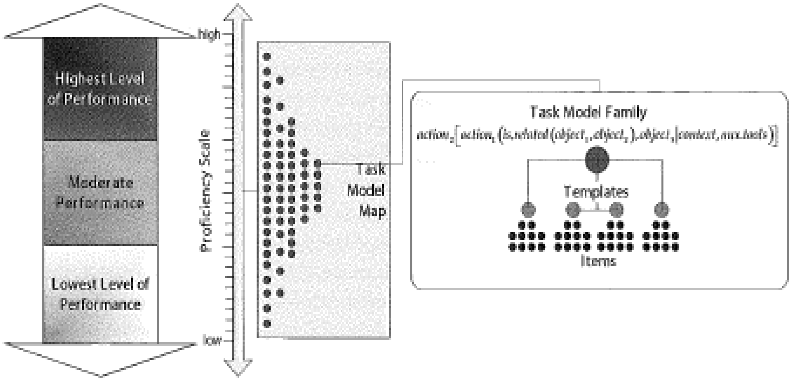
Figure 1. Layers of evidence-centered design.

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. Military Medicine, 178, (Supplement 1), 107-114.

AIG in practice

- Weak and strong theory approaches to item generation (Drasgow, Luecht, & Bennett, 2006)
- The expedient bottom-up approach
  - Use an existing item pool from which to generate templates or item models that are the basis for generation (Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003)
- A principled top-down approach
  - Hendrickson, Huff, Luecht (2010)
  - Luecht (2012)

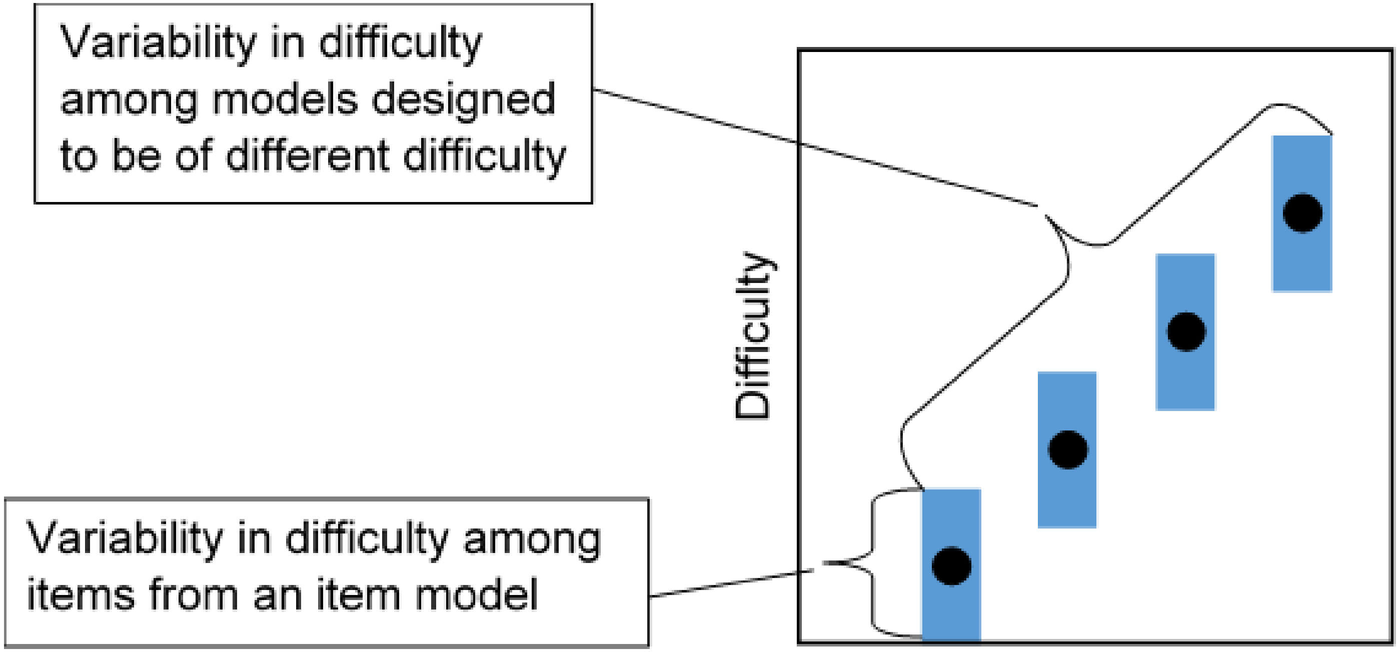



Highest Level of Performance
Moderate Performance
Lowest Level of Performance
high
low
Proficiency Scale
Task Model Map
Task Model Family
action₂[action₁(is.related(object₁, object₂), object₃ | context, aux.tools)]
Templates
Items

Verbal difficulty modeling: a sobering reality

Source	Items (N)	Method	Results: R^2	Comments
(Bejar, Deane, Flor, & Chen, 2017)	GRE® Sentence equivalence (800)	Linear regression	.12	Operational GRE® items Context + familiarity + depth of familiarity
(Deane, Lawless, Li, Sabatini, Bejar, & O'Reilly, 2011)	Vocabulary	Linear regression	.41	Experimental items Similarity between target, keys and distractors as measured by the cosines of semantic vector word representations, word frequency
(Embretson & Wetzel, 1987)	Paragraph comprehension	Linear regression	.29	Operational ASVAB items Text characteristics + decision processes
(Gorin & Embretson, 2006)	Paragraph comprehension	Linear regression	.34	Operational GRE® items TR + DP + GRE-specific factors
(Enright & Bejar, 1989)	Analogies	Regression + human judgment	.43, .44	Operational GRE® items Semantics class + rationale complexity
(Carroll,1980)	SAT® antonyms	Correlation	.58	Illustrative results Standardized frequency
Sheehan & Mislevy				

Growing number of psychometric models


Variability in difficulty among models designed to be of different difficulty

Variability in difficulty among items from an item model

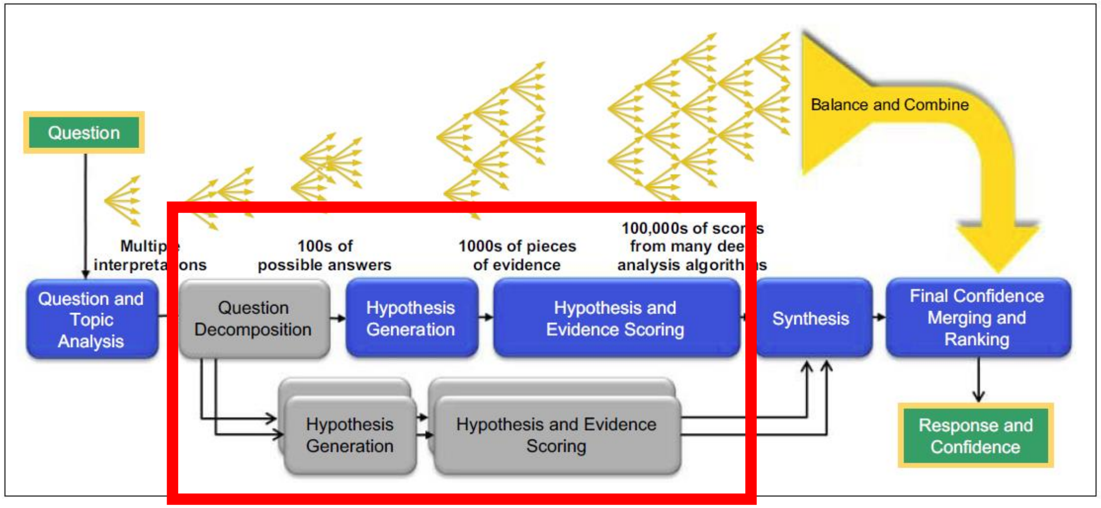
Fischer, G. H. (1973). The linear logistic model as an instrument of educational research. *Acta Psychologica, 37,* 359-374.
De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73(4),* 533-559.
Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika, 64,* 407-433.

Summary and conclusions

Answer modeling

Question answering

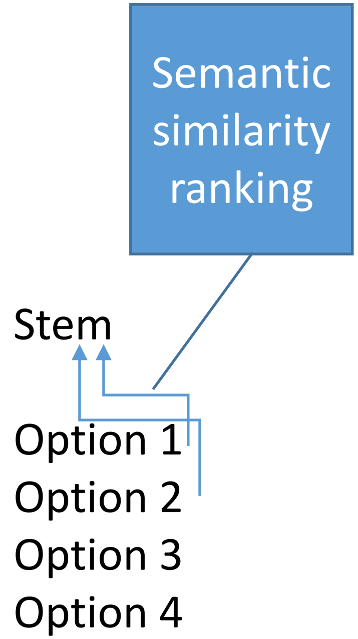
- Using multiple sources of information identify potential answers, rank them, and choose the top ranked answer.
  - (See Jurafsky, D., & Martin, J. H. (2016). Question answering. In Speech and language processing Retrieved from: https://web.stanford.edu/~jurafsky/slp3/28.pdf:)
- IBM Watson wins at Jeopardy! (And it is now 240 times “stronger”)



The Science And Technology Behind IBM Watson (Part 2 of 5 Series)

TOEFL vocabulary (State of the art)

Source	Approach	Performance	Reference
Tsatsaronis et al. (2010)	Lexicon-based	87.50%	Tsatsaronis, G., Varlamis, I., and Vazirgiannis, M. (2010). Text Relatedness Based on a Word Thesaurus. Journal of Artificial Intelligence Research 37, 1-39
Dobó and Csirik (2013)	Corpus-based	88.75%	Dobó, A., and Csirik, J. (2013). Computing semantic similarity using large static corpora. In: van Emde Boas, P. et al. (eds.) SOFSEM 2013: Theory and Practice of Computer Science. LNCS, Vol. 7741. Springer-Verlag, Berlin Heidelberg, pp. 491-502
Rapp (2003)	Corpus-based	92.50%	Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. Proceedings of the Ninth Machine Translation Summit, pp. 315-322.
Pilehvar et al. (2013)	WordNet graph-based (unsupervised)	96.25%	Pilehvar, M.T., Jurgens D., and Navigli R. (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria.
Turney et al. (2003)	Hybrid	97.50%	Turney, P.D., Littman, M.L., Bigham, J., and Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03), Borovets, Bulgaria, pp. 482-489.
Bullinaria and Levy (2012)	Corpus-based	100%	Bullinaria, J.A., and Levy, J.P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. Behavior Research Methods, 44(3):890-907.



Semantic similarity ranking

Stem

Option 1

Option 2

Option 3

Option 4

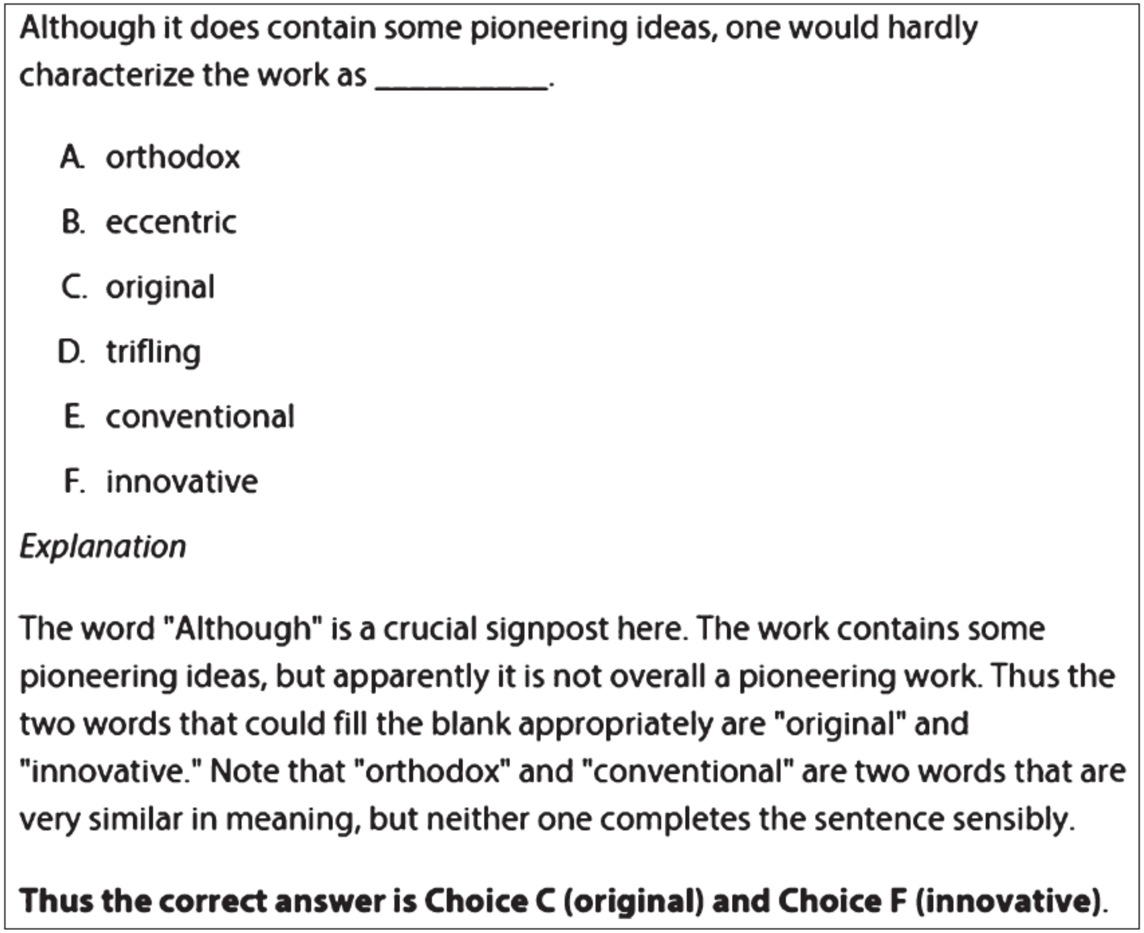
SAT analogies (State of the art)

Source	Approach	Performance	Reference
Turney and Littman (2005)	Human	57.0%	Turney, P.D., and Littman, M.L. (2005). Corpus-based learning of analogies and semantic relations. Machine Learning, 60 (1-3), 251-278.
Turney (2008)	Corpus-based	52.1%	Turney, P.D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, pp. 905-912.
Turney (2006a)	Corpus-based	53.5%	Turney, P.D. (2006a). Expressing implicit semantic relations without supervision. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling/ACL-06), Sydney,
Turney (2013)	Corpus-based	54.8%	Turney, P.D. (2013), Distributional semantics beyond words: Supervised learning of analogy and paraphrase, Transactions of the Association for Computational Linguistics (TACL), 1, 353-366.
Speer et al. (2017)	Hybrid	56.1%	Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. Proceedings of The 31st AAAI Conference on Artificial Intelligence, San Francisco, CA.
Turney (2006b)	Corpus-based	56.1%	Turney, P.D. (2006b). Similarity of semantic relations. Computational Linguistics, 32 (3), 379-416.

AI answer models not necessarily based on a cognitive model

- The goal of answer models is to answer *all* questions
- Models could be modified to be more cognitively oriented
  - They share elements in common with information processing models
- The methods behind answer modeling could have uses in assessment design
  - Item critique
  - Item analysis
  - Feature extraction for difficulty modeling (CR)
  - Detect effective gaming strategies (CP)

Applications to GRE® verbal items



Although it does contain some pioneering ideas, one would hardly characterize the work as ___________.

A. orthodox

B. eccentric

C. original

D. trifling

E. conventional

F. innovative

Explanation

The word "Although" is a crucial signpost here. The work contains some pioneering ideas, but apparently it is not overall a pioneering work. Thus the two words that could fill the blank appropriately are "original" and "innovative." Note that "orthodox" and "conventional" are two words that are very similar in meaning, but neither one completes the sentence sensibly.

Thus the correct answer is Choice C (original) and Choice F (innovative).

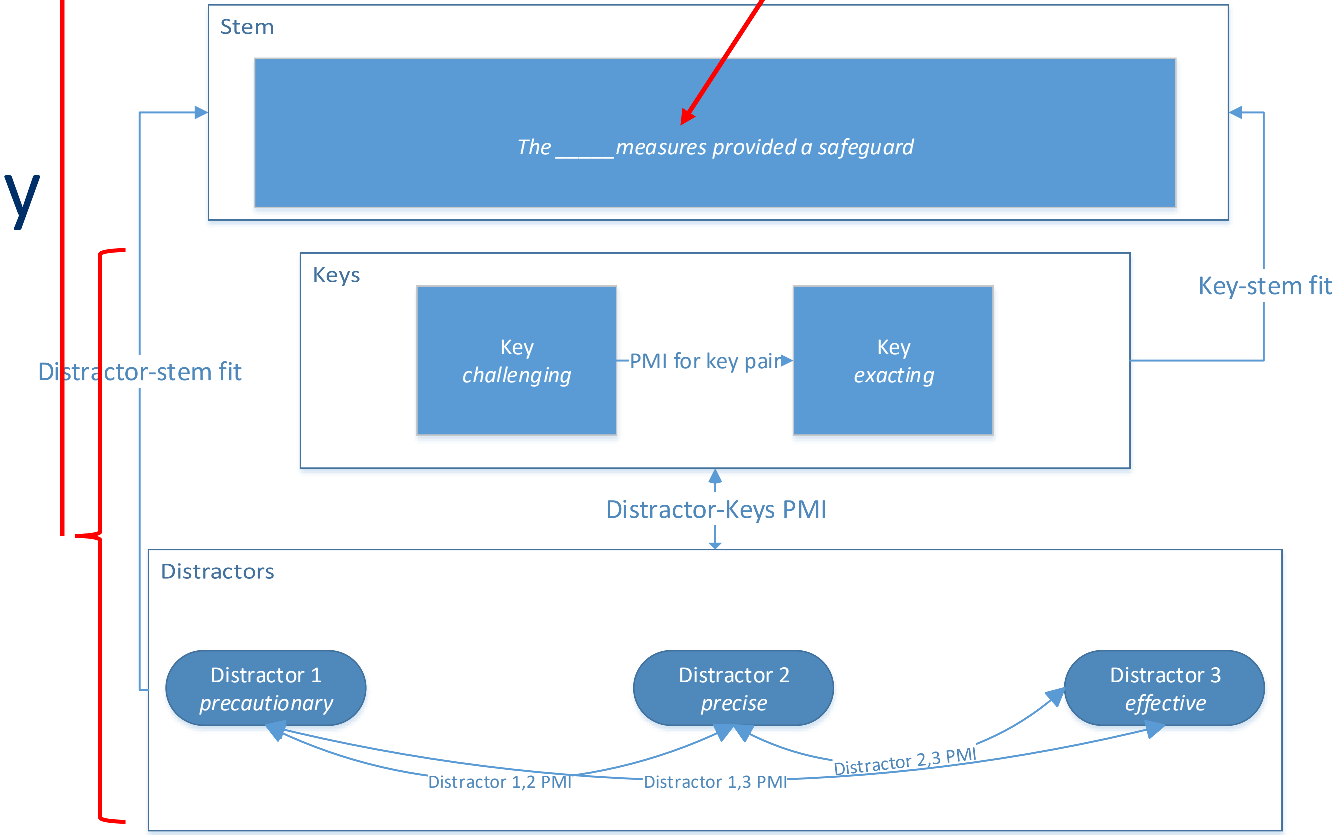
Figure 1 Sample sentence equivalence item type (adjective).

3) Depth of familiarity

1) Context

2)Familiarity

Challenging
Exacting
Precautionary
Precise
Effective



GRE sentence equivalence item type: Not just vocabulary

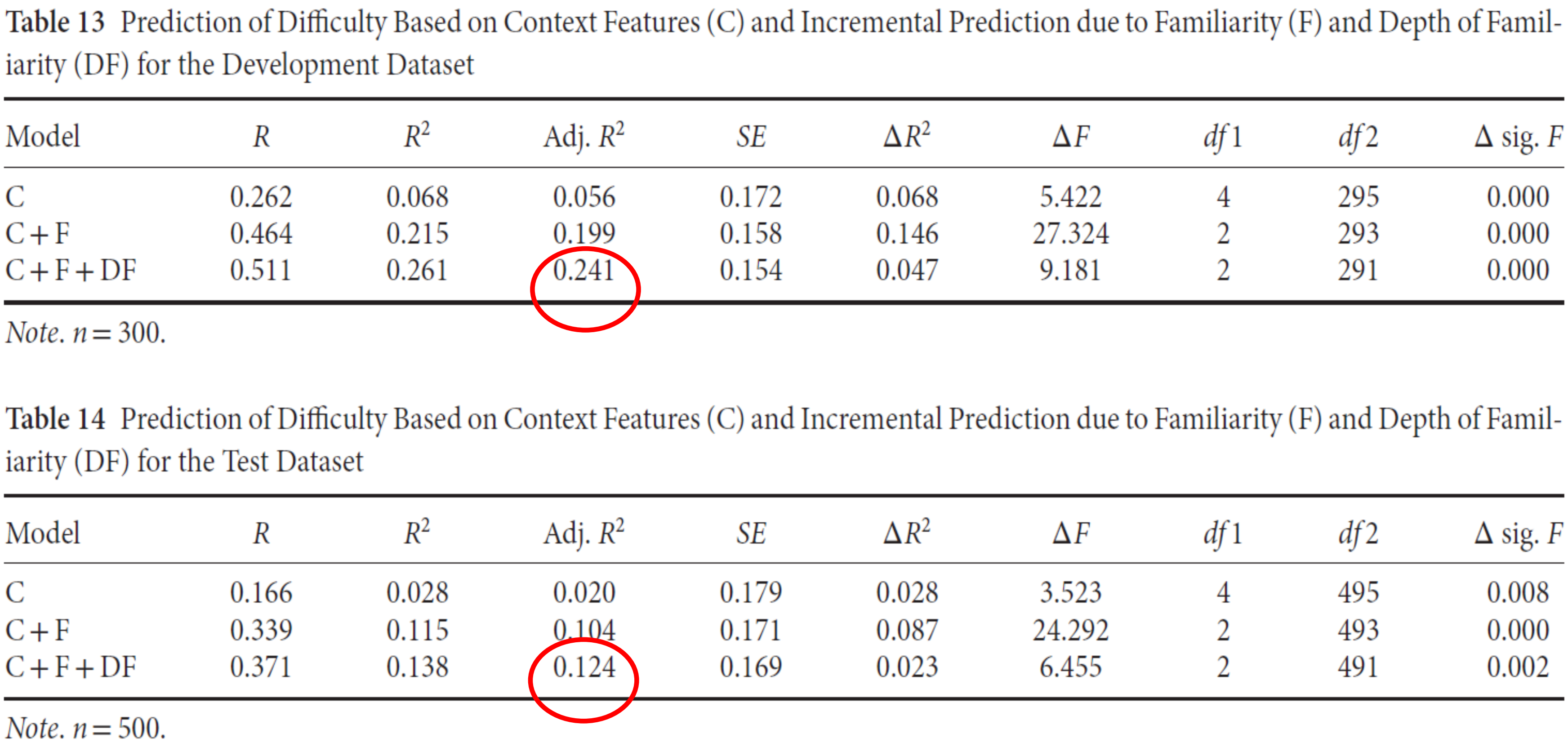


Table 13 Prediction of Difficulty Based on Context Features (C) and Incremental Prediction due to Familiarity (F) and Depth of Familiarity (DF) for the Development Dataset

Model	R	R^2	Adj. R^2	SE	ΔR^2	ΔF	$df1$	$df2$	Δ sig. F
C	0.262	0.068	0.056	0.172	0.068	5.422	4	295	0.000
C + F	0.464	0.215	0.199	0.158	0.146	27.324	2	293	0.000
C + F + DF	0.511	0.261	0.241	0.154	0.047	9.181	2	291	0.000

Note. $n = 300$.

Table 14 Prediction of Difficulty Based on Context Features (C) and Incremental Prediction due to Familiarity (F) and Depth of Familiarity (DF) for the Test Dataset

Model	R	R^2	Adj. R^2	SE	ΔR^2	ΔF	$df1$	$df2$	Δ sig. F
C	0.166	0.028	0.020	0.179	0.028	3.523	4	495	0.008
C + F	0.339	0.115	0.104	0.171	0.087	24.292	2	493	0.000
C + F + DF	0.371	0.138	0.124	0.169	0.023	6.455	2	491	0.002

Note. $n = 500$.

Bejar, I. I., Deane, P., Flor, M., & Chen, J. (2017). *Evidence of the generalization and construct representation inferences for the GRE® sentence equivalence item type.* ETS GRE-17-02, ETS RR-1705, Princeton, NJ: ETS.

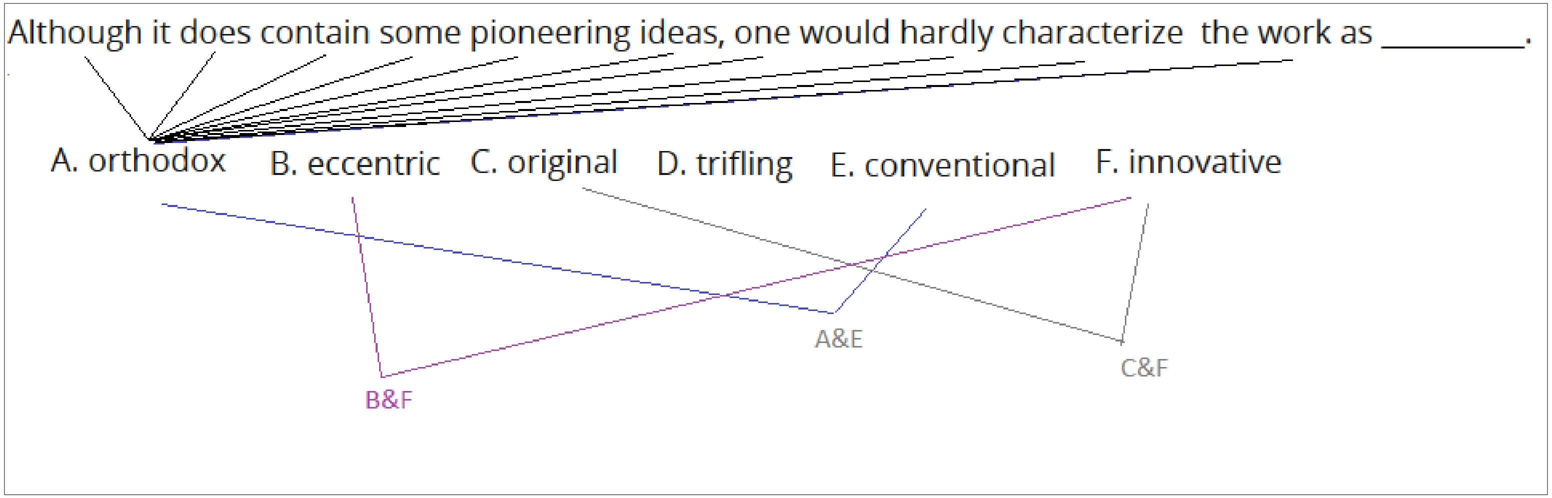
Answer modeling applied to GRE® sentence equivalence item type

- Can an answer model solve the items?
    - Nature of items it solves?
- Is the answer modeling process useful in difficulty modeling?
- Approach
    - Represent words as embeddings (word2vec)
    - Compute similarity among key and distractors as the cosine of their vectors:

$$cosine\ (\boldsymbol{a}, \boldsymbol{b}) = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}}$$

Cosine of france with

spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130
switzerland	0.622323
luxembourg	0.610033
portugal	0.577154
russia	0.571507
germany	0.563291



Although it does contain some pioneering ideas, one would hardly characterize the work as ________.

A. orthodox B. eccentric C. original D. trifling E. conventional F. innovative

1. Contextual Fit: for each response option, compute average cosine with all content words in the stem

2. Inter-Option Relatedness: for each pair of response options,
compute their cosine similarity (trying to capture synonymy)

	1. Contextual Fit: for each response option, compute average cosine with all content words in the stem. 1. Inter-Option Relatedness: for each pair of response options, compute their cosine similarity (trying to capture synonymy)	Word embeddings: Context: w_1 w_2 w_3 w_4..w_{nw} Distractors: D_j j=1..6 Where, $Sim(D_j,D_m) = \cos(D_j , D_m)$, $j \neq m$, $i=1..6$, $m=1..6$
Synonymy	Compute 15 pairwise inter-option similarities, rank the pairs and pick the one with highest similarity	$Max(Sim(D_j,D_m))$
Best context	Compute contextual fit for each option, rank, pick the two options with highest contextual fit.	Sim(D_j,Context) =($\sum_i^{nw} cos$(D_j,W_i))/nw Choose top 2.
Best context, then synonym	Compute contextual fit for each option, rank, pick the one option with highest contextual fit, then find best synonym for that option.	
Best contexts and synonym	For each pair, compute inter-option similarity, add contextual fit of both options. Rank pairs, pick best ranking.	

Results (300 sentence equivalence GRE® items)

	Method (strategy)	Wordnet	word2vec	Both
1	Best context		48	48
2	Synonymy	49	109	125
3	Best context, then synonym		66	83
4	Best contexts and synonym		109	**129**

Notes:

Detection of synonymy is by far the most important contributor.

Best results (43% of the items correctly solved), were obtained by combining synonymy-detection with some contextual fit.

Average difficulty of solved items is -0.10
Average difficulty of unsolved items is 0.17

Adding to difficulty modeling

- Poor relative ranking for the key pair is weakly correlated with item difficulty
  - r=0.15 with 300 sentence equivalence GRE items

Generation of sentence equivalence items

- Given a stem:
- Find six words (K1,K2, D1, D2, D3, D4) from VOCABULARY such that,
  - Context constraints:
    - FIT(K1)-FIT(K2) is…
    - FIT(D1), FIT(D2),FIT(D3),FIT(D4) is…
  - Familiarity constraints
    - SFI(K2) <= SFI(K1), SFI(D_i) <= SFI(K1)
  - Depth of familiarity constraints
    - SIM(K1,K2) is…
    - SIM(D1,D2,D3,D4) is..

The *exacting* __ precautions were a safeguard

K1

K2

D1

D2

D3

D4

Summary and conclusions