IMPLEMENTING ARTIFICIAL INTELLIGENCE SOLUTIONS FOR COMMON TEST DEVELOPMENT CHALLENGES

THE TWENTIETH ANNUAL MARC CONFERENCE: MACHINE LEARNING, NATURAL LANGUAGE PROCESSING, AND PSYCHOMETRICS

KIMBERLY SWYGERT
IAN MICIR
NBME
NOVEMBER, 2022

WHO WE ARE

- Kimberly Swygert, PhD, Director, Test Development Innovations, NBME
  - https://www.linkedin.com/in/kimberlyswygert/
- Ian Micir, BA, Designer, Test Development Innovations, NBME
  - https://www.linkedin.com/in/ian-micir-77a55864/

WHAT NBME DOES

- Protects the health of the public through state-of-the-art assessment
- Flagship product: USMLE series (Steps 1-3) for initial physician licensure
- Massive volume = extensive item bank review and maintenance
- Test Development: ~100 staff, extensive item development and review processes, support vigilant test security
- State of the art is a state of mind

AI IN ASSESSMENT

- AI within educational fields, including medical education, is moving forward at a rapid pace, where AI tools are being deployed within the context of learning and tutoring – opening “the black box of learning”
- AI within assessment, including licensure assessment, is being implemented more slowly than the educational or field of practice uses of AI
- Recently, there have been promising indicators:
  - *Principles of AI Use In Testing* (Association of Test Publishers, 2021)
  - *Computational Psychometrics* (von Davier, Mislevy, & Hao, 2022)
  - Pandemic response necessitating AI tools, such as AI proctoring

AI IN ASSESSMENT

- However, the challenges are substantial, with technology and data not always the biggest challenges
- Test developers abide by the *Standards,* while the use of AI introduces a new set of ethical standards that overlap but are not identical
- Regulation of data use for AI may be in addition to regulations for responsible governance and use of data within testing organizations
- Profound ethical and human-centered challenges are being posed anywhere AI is used, even if not for assessment

AI AT NBME

- We began organically and “bottom-up” - What existing item and test development challenges could be addressed by allowing our business expertise to guide applications, developed in-house, that utilized NLP and ML?
- Other testing organizations might be asking similar questions:
  - How can AI support item coding and pool management when non-NLP automated item generation is in use?
  - How can AI protect test security via rapid matching?
  - How can AI provide better searches for SMEs within large item or multimedia storage systems?
  - How can AI help us identify stereotyped or problematic language patterns within item banks?

AI AT NBME

- In alignment with this, we focused on:
  - Coordination with AI research efforts and other business uses of AI
  - Coordination with internal data strategy, privacy, and governance efforts
  - Scanning the landscape for AI use in medical education and practice
  - Developing of an AI Center of Excellence
  - Planning for organizational strategy that would guide future efforts
- We also realized that crucial ethical concerns may emerge, even with AI development on a small scale where examinees were not directly interacting with the AI

ETHICAL CAUTIONS AND OUR AI PRINCIPLES

Most people who bother with the matter at all would admit that the English language is in a bad way, but it is generally assumed that we cannot by conscious action do anything about it. Our civilization is decadent and our language – so the argument runs – must inevitably share in the general collapse. It follows that any struggle against the abuse of language is a sentimental archaism, like preferring candles to electric light or hansom cabs to aeroplanes. Underneath this lies the half-conscious belief that language is a natural growth and not an instrument which we shape for our own purposes.

Now, it is clear that the decline of a language must ultimately have political and economic causes: it is not due simply to the bad influence of this or that individual writer. But **an effect can become a cause, reinforcing the original cause and producing the same effect in an intensified form, and so on indefinitely**. A man may take to drink because he feels himself to be a failure, and then fail all the more completely because he drinks. It is rather the same thing that is happening to the English language. It becomes ugly and inaccurate because our thoughts are foolish, but the slovenliness of our language makes it easier for us to have foolish thoughts. The point is that the process is reversible. Modern English, especially written English, is **full of bad habits which spread by imitation and which can be avoided if one is willing to take the necessary trouble**. If one gets rid of these habits one can think more clearly, and to think clearly is a necessary first step toward political regeneration: so that **the fight against bad English is not frivolous and is not the exclusive concern of professional writers.**

-George Orwell, *Politics and the English Language*

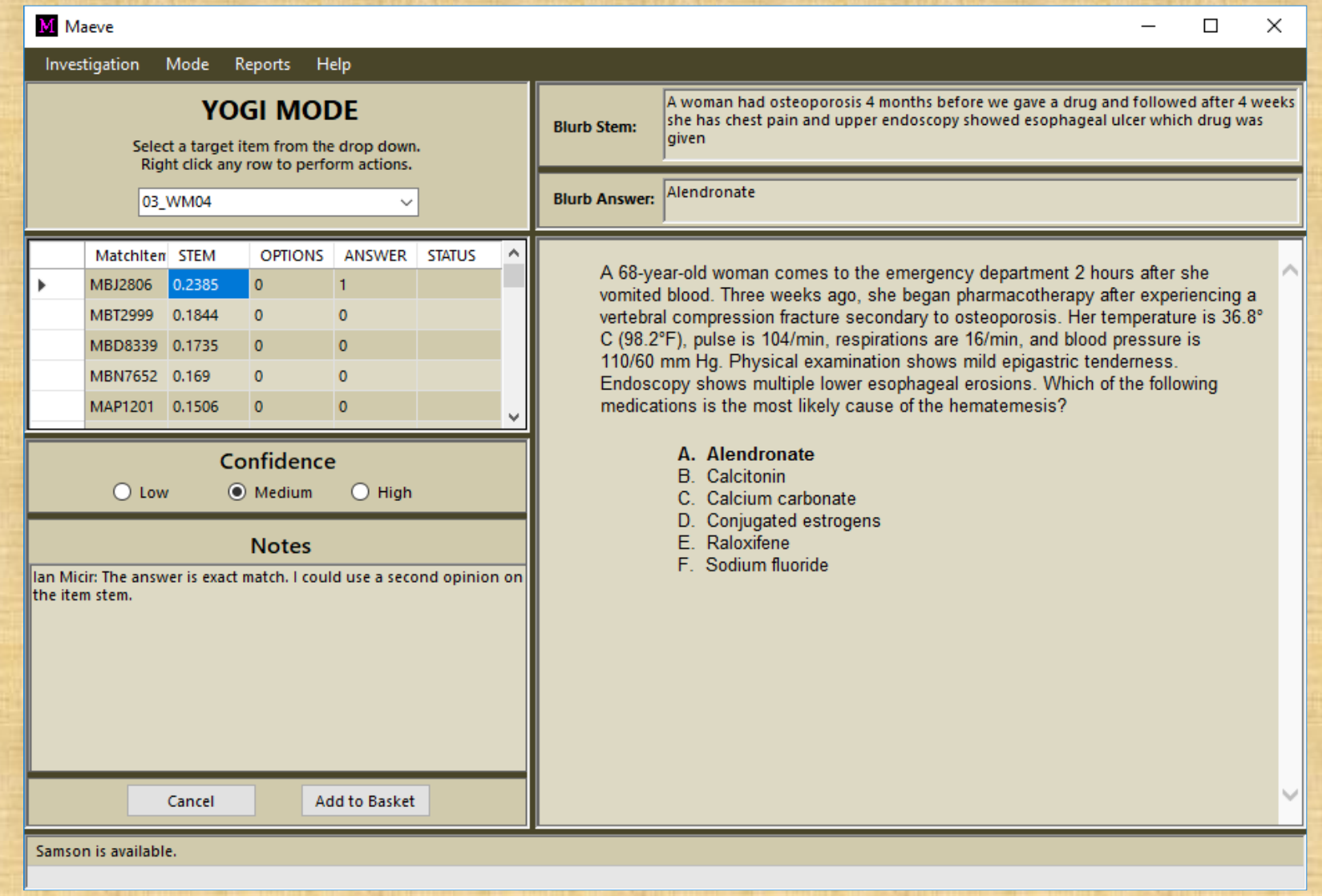
ETHICAL CAUTIONS AND OUR AI PRINCIPLES

- “Move fast and break things” vs. “Paralysis from fear of mistakes”
- Easy to get lost in the coolness of it all… because it *is* cool!
- In order: *Should we do it? Can do we it? How should we do it?*
- Moral responsibility and respect for human dignity
- Let’s learn from ML – stop saying “human in the loop”
- Our goal: accuracy, efficiency, and employee satisfaction

OUR AI APPLICATIONS – YOGI

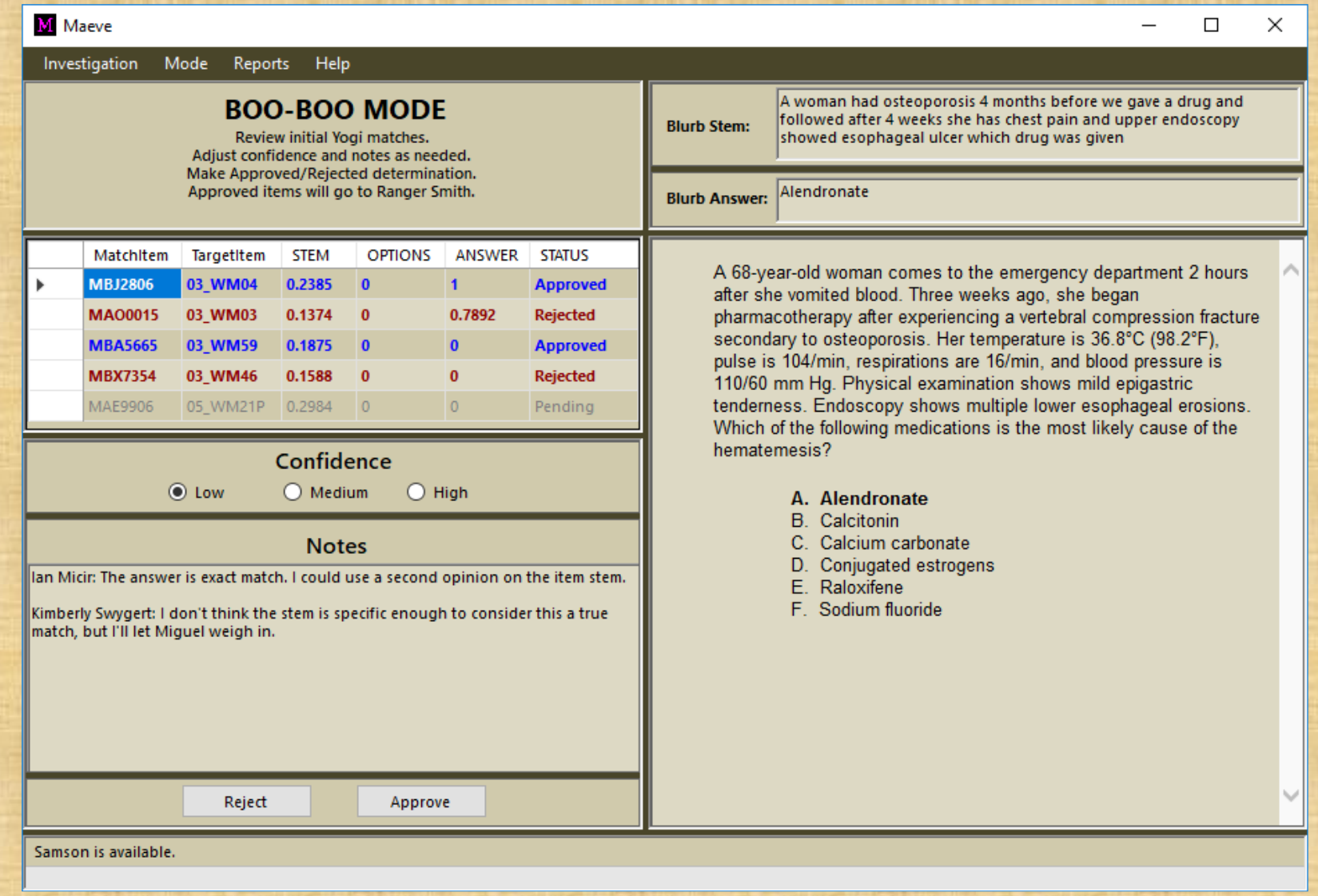
- .NET application written in C#
- Compares blurbs found on the internet to our item banks
- Automates a good deal of pre-editing via Regex patterns
- Uses TF-IDF vectors to calculate cosine similarity (stems & answers)
- Metrics: Previous method vs. Yogi
- Slides from 2019 presentation at Conference on Test Security (COTS) available upon request

OUR AI APPLICATIONS – YOGI



Previously presented at COTS, 2019

OUR AI APPLICATIONS – YOGI



Previously presented at COTS, 2019

OUR AI APPLICATIONS – SMOKEY

- .NET application written in C#
- Extension of Yogi NLP engine with ML element added
- Parameters: stem & answer similarity, *n* content codes, enemy status
- Supervised learning: existing enemy status = truth variable
- Metrics included editor satisfaction and reduction of hours needed to complete the tasks in addition to accuracy
- Published in the *Journal of Applied Testing in Technology*, 2002
- https://www.jattjournal.com/index.php/atp/issue/view/9063

OUR AI APPLICATIONS – SMOKEY

FALSE POSITIVES					
RUN	EXAM ENEMY	SUBFORM ENEMY	NOT ENEMY	REVIEW	TOTAL
1	100	0	0	0	100
2	73	27	0	0	100
3	84	16	0	0	100
4	91	9	0	0	100
5	97	3	0	0	100
6	79	21	0	0	100
7	66	0	23	11	100
TOTAL	590	76	23	11	700

FALSE NEGATIVES					
RUN	EXAM ENEMY	SUBFORM ENEMY	NOT ENEMY	REVIEW	TOTAL
1	7	3	14	1	25
2	7	6	10	0	23
3	12	22	12	0	46
4	33	15	4	0	52
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
TOTAL	59	46	40	1	146

OUR AI APPLICATIONS – SECRET TO SUCCESS

- Intuitive UI: users focus on the work, not operating the program
- A great idea is not 90% of the work
- A successful research experiment is only the *beginning* of the process
- Staff empowerment: our *editors* determine the model parameters
- The *program* is in the loop, not the person

LESSONS LEARNED

- Define the small projects where AI can “allow humans to be the best experts they can be,” with accompanying metrics
- AI efforts should be coordinated with strategy, data governance, research, and change management plans to educate and support staff
- Design specialized training for SMEs and non-technical staff, and explicitly address concerns from staff about their skillsets and the technical demands of innovation
- Deploy effective and targeted communication that can counteract more fantastical examples of AI in fictional media

Kimberly Swygert, kswygert@nbme.org
Ian Micir, imicir@nbme.org

THANK YOU!