

Funnel Plot Analysis of Examinee Performance in Longitudinal Assessment

Aquia Richburg¹, Marcus Walker²

¹aquiar@nccpa.net, ²marcusw@nccpa.net

National Commission on Certification of Physician Assistants (NCCPA)

Abstract

We identify content areas where examinees underperform in a longitudinal assessment for a high-stakes recertification exam. Using modified funnel plots with a moving baseline to account for item difficulty, we rank topics by relative performance. Preliminary results suggest reasons beyond item difficulty, informing future analyses supporting potential educational interventions.

1 Introduction

Research has shown that self-testing one's own knowledge improves long-term retention over simply restudying material in what is known as the testing effect (Roediger and Karpicke, 2006). Reflecting this principle, healthcare certification programs have increasingly adopted longitudinal assessment (LA) as an alternative to single-point-in-time examinations, with studies suggesting that LA supports continuous learning, identifies knowledge gaps, and improves performance while maintaining assessment of competence (Fraundorf et al., 2023; Giron et al., 2021; Ward et al., 2023). As LA becomes more widely implemented, understanding patterns of examinee performance across content domains may provide valuable insights for optimizing lifelong learning and informing assessment design.

2 Background

Analyzing the change in student performance over time is an important component in measurement theory. Understanding the trajectories of different subpopulations provides insights informing decisions on pedagogical and testing designs. One application of this analysis is the identification of knowledge gaps. These can be thought of as the discrepancy of an individual's ability in relation to an expected performance threshold and the reasons

for these gaps can come in many forms. Schmidt (2020) outlines an extensive taxonomy of knowledge gaps and potential strategies to address them. This taxonomy ranges from more surface level, such as understanding textual content, to more abstract, such as requiring implicit context to arrive at a conclusion. In the field of medical licensure, the more metacognitive categories of knowledge gaps could be a better theory for explaining performance difficulties due to the multilayer complexity of medical-domain exam items.

Research in measuring growth trajectories has taken a variety of different forms. Early works like Muthen and Khoo (1998) generalize random coefficient models to incorporate latent variables that can more directly model individual variation in growth parameters between multiple subpopulations. Latent variable frameworks have continued to be popular in recent research where various methods have been proposed to measure and predict population-dependent growth trajectories and in some cases to identify points for intervention (von Davier et al., 2011; Kwon and Zhang, 2025; Liu and Culpepper, 2025; Yamaguchi et al., 2026). While these prior studies primarily use repeated measurements of performance over a longitudinal period, our work differs by also incorporating longitudinal performance of items administered to address deficits in knowledge.

Natural language processing (NLP) has become a powerful tool in the biomedical domain. To address the distributional shift in text from a general domain, several language models (Alsentzer et al., 2019; Lee et al., 2020; Rasmy et al., 2021) have been fine-tuned for tasks such as information extraction from domain-specific or structured text. In more recent work, large language models (LLMs) have been integrated into systems to reformulate unstructured biomedical

textual data into formats more readily usable (Wiest et al., 2025; Hsu and Roberts, 2025). Item difficulty prediction has also become a task that increasingly relies on NLP techniques. Contreras (2018) present methods to measure the relative frequency in terms between topic-specific and general texts in an attempt to quantify differences in mental workload among examinees. Ha et al. (2019) construct a set of linguistic and information retrieval features for item difficulty prediction using a corpus constructed from the United States Medical Licensing Examination® (USMLE®). The BEA (Building Educational Applications) 2024 shared task (Shardlow et al., 2024) focuses on automatic difficulty prediction and response time prediction using the USMLE® where the best strategy relied on transformer-based text embeddings. Li et al. (2025) achieved a slightly better model than the best shared task participant by incorporating data augmentation alongside language models.

3 Aims

The main goal of this study is identifying global patterns on a longitudinal assessment using the response data over a cohort of examinees. The exam blueprint consists of several main content areas that are comprised of various topics. During each window of administration, an examinee is presented with a set of items representatively sampled from the distribution of topics. Over the course of the examination, the set of items changes according to response behavior, favoring more items from topics an examinee missed. Identifying topic areas of the blueprint where examinees are more likely to have difficulty than expected could better support the goal of continuous learning.

4 Data

We used the response data for a longitudinal exam program consisting of 12 administration windows (quarters) from 2023 to 2025. Each quarter examinees are provided with 25 items across 14 content areas that can target one of four cognitive tasks. Examinee responses to items are categorized as an original attempt, items drawn from a pool according to blueprint specifications, or a targeted attempt, items chosen according to blueprint specifications **and** addressing a topic encountered in a prior quarter according to an index that incorporates whether an item was answered

incorrectly, the confidence of an examinee on that topic and the examinee's perceived relevance of that topic to their clinical practice. Higher index scores increase the likelihood that the topic will appear in future quarters. Each examinee is tested over 12 quarters and may exit early by passing 8 quarters. We hypothesize that low performance on targeted attempts relative to original attempts could signify gaps in an examinee's knowledge on specific topics.

5 Methods

5.1 Confidence and Targeted rankings

The structure of this longitudinal exam is semi-adaptive; the items administered to examinees is informed by information from prior windows of information. After the completion of each item, an examinee is asked to rate their confidence in answering the item and rate the perceived relevance of the item to their practice where higher values indicate stronger agreement. In the attempt to serve the most useful items to examinees, a targeted index is computed incorporating confidence, relevance and the binary performance on the item. The exam prioritizes selecting subsequent items from topics where examinees incorrectly answer the item and have high targeted index.

Confidence ratings and targeted index scores provide direct signals from examinees interacting with the exam; thus ranking items along these dimensions could indicate areas where examinees struggle. For each topic, we compute the mean confidence rating and mean targeted index across all responses from 2023 to 2025. We rank confidence from lowest to highest and rank targeted index from highest to lowest. These rankings are included as comparisons against our funnel ranking which we will describe in the next section.

5.2 Funnel ranking

The main methodology we incorporate is the statistical funnel plot. A funnel plot allows for the comparison of different indicators where the precision of the indicator varies. Spiegelhalter (Spiegelhalter, 2005; Spiegelhalter et al., 2012) utilized funnel plots to propose an alternative to rank medical institutions. Instead of comparing performance indicators using a common confidence interval, they accommodate differences

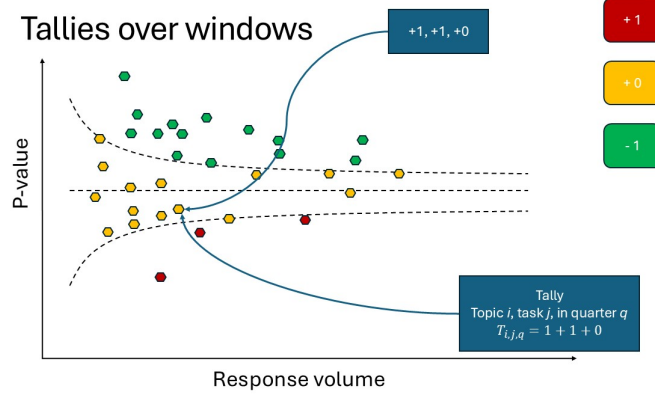


Figure 1: Graphic illustrating the tallying of scores across funnel plot windows. This example demonstrates the tally accrued over three downward consecutive funnel plot windows in funnel window $n = 3$.

in precision by varying the confidence interval to allow for differences in indicator variance. In turn, this creates a funnel-like shape around a chosen baseline (which could be the mean of the indicator) as the confidence interval shrinks, where precision increases. Points that fall within the control limits are not statistically different from each other while those outside have reason for scrutiny.

Let $a_{i,j,q,h}$ be an item attempt that consists of a topic i and task j that is administered in quarter q to examinee h . We define the P-value of a topic-task-quarter combination, $p_{i,j,q}$, as the ratio of all correct responses to incorrect responses for attempts $a_{t,j,q,h}$ across all examinees. We assign a mean difficulty, $\mu_{b_{i,j,q}}$, for the combination of topic-task-quarter, $p_{i,j,q}$, by computing the mean item difficulty parameter $\mathbf{b}$ from the Rasch model across the items of topic i and task j in quarter q . Thus, our points of interest are topic-task combinations across quarters, and the indicator is the associated P-value. Since performance on items is dependent on its difficulty, a common baseline P-value over all items would disadvantage items towards the tail ends of the distribution, since they are further from the mean.

Let A_O and A_T be the collection of $p_{i,j,q}$ computed from original attempts and targeted attempts, respectively. Using the distribution of mean difficulties, $\mu_{b_{i,j,q}}$, we first partition the collection of all type-task-quarter combinations ($p_{i,j,q} \in A_O \cup A_T$) using k -means. This generates N classes of $p_{i,j,q} \in A_O$ for comparisons against $p_{i,j,q} \in A_T$. For our study, we set $N = 5$. Within the structure of the longitudinal exam, examinees usually enter in the first quarter of a calendar year generating a cohort that is administered items from

a common pool. In turn, targeted attempts are pulled from the same common pool for subsequent quarters. For example, the 2023 cohort will see original attempt items from the 2023 pool in quarters 1 through 4; then the same cohort will see targeted attempt items from the same pool in quarters 2 through 5 (quarter 5 is equivalent to quarter 1 of 2024). To keep comparisons of items consistent within a pool we partition both A_O and A_T into years $y = 1, 2, 3$ for 2023, 2024 and 2025, respectively. Since targeted attempts are administered after original attempts, the years are staggered by a quarter, i.e. $y = 1$ consists for quarters 1-4 for A_O and 2-5 for A_T . For each partition, we define a baseline P-value as the median of the mean difficulties, $\mu_{b_{i,j,q}}$, from $p_{i,j,q} \in A_O^y$. We choose the median value to avoid possibly skewed distributions of mean difficulties. The baseline P-values are then used as points of comparison against $p_{i,j,q}$ from the corresponding A_T^y .

We plot $p_{i,j,q} \in A_T^y$ for each $y \in (1, 2, 3)$ where the volume of responses and the P-value represent the x - and y -axis, respectively. Using the baselines defined above, we plot $n = N$ funnel plots overlapping the data. For each type-task-quarter combination, we tally scores, $T_{i,j,q,n}$, by assigning -1, 0 or 1 according to whether the point $p_{i,j,q}$ lies above, within or below the control limits for each funnel plot, respectively. We then sum the scores $T_{i,j,q,n}$ across all N funnel plots for each quarter q . This generates a score for each topic-task-quarter combination. See figure 1 for a graphic depicting the process.

We compute final scores, $S_{i,j}$, for topic-task combinations as the mean $T_{i,j,q}$ over all quarters;

and for topics, S_i , as the mean score $S_{i,j}$ over tasks in topic i . Ultimately, we rank all S_i or $S_{i,j}$ from highest to lowest where higher scores indicate areas where examinees tend to have more difficulty. We allow ties to prevent spurious rankings. The sequence of ranking equations are listed below:

$$T_{i,j,q} = \sum_{n=1}^N T_{i,j,q,n}, \quad (1)$$

$$S_{i,j} = \frac{1}{Q} \sum_{q=1}^Q T_{i,j,q}, \quad (2)$$

$$S_i = \frac{1}{m_j} \sum_{m=1}^{m_j} S_{i,j} \quad (3)$$

where N is the number of funnel plot windows, Q is the total number of quarters and $m_j \in [1, 4]$ is the number of tasks in topic i .

6 Results

Within each of the 14 content categories, we list the top ranked topics in Table 1 along with its allocation according to the blueprint¹. Unsurprisingly, the highest-ranked topic is from the cardiovascular system, given its allocation is the highest at 12%. The remaining ranks do not appear to follow a pattern; the genitourinary system is only 5% but has a topic at rank 3 while gastrointestinal and pulmonary (both 10%) have ranks at 13 and 18, respectively. Similarly, item difficulty does not have a clear connection to rank; compare rank 1 with difficulty 1.262 and rank 18 with difficulty 2.038. We also include the overall top 10 funnel-ranked topics in Table 3 and the highest- and lowest-ranked topics for each ranking type in Table 2.

In figure 2, we plot the correlation of funnel, confidence and index rankings against mean topic difficulty. All three plots show a negative correlation against item difficulty; however confidence and index rankings have tighter groupings of items across rank levels compared to the funnel rankings. Quantitatively, we compute the Kendall τ correlation as -0.416, -0.566 and -0.705 for funnel, confidence and index rankings, respectively. This indicates that the funnel

rankings describe trends in examinee behavior that are not explained by item difficulty.

For the topic with the highest mean difficulty (hypertension, secondary in Cardiovascular System with difficulty 3.527), there is a large difference between the funnel rank at 80 and confidence and index ranks at 15 and 17, respectively. In contrast the lowest mean difficulty topic (rabies in Infectious Diseases with difficulty -3.238) has a slightly smaller difference between ranks with 181, 237 and 237 for funnel, confidence and index ranks, respectively. The greatest difference in funnel and index ranks is $225 - 82 = 143$ for pneumonia: bacterial, fungal, viral (Pulmonary System) and the greatest difference in funnel and confidence ranks is $212 - 31 = 181$ for valvular disorders: regurgitation, prolapse, stenosis (Cardiovascular System). We also compute Kendall τ as 0.381 for funnel against confidence ranks, 0.497 for funnel against index ranks and 0.608 for confidence against index ranks.

7 Discussion and Conclusion

Our main goal for this study was to use the response data from a high-stakes longitudinal recertification exam to identify topics where examinees particularly struggle. Within the structure of the exam, there are survey questions that ask about each item's relevance to the examinee and the confidence in their answer. Although these self-reported scores provide a signal for identifying topics that might be more difficult than others, there is a risk of bias that might distort the actual difficulty seen in the data. By utilizing the more objective response data, we can more directly measure the variation in performance across different areas of the blueprint.

Our results show there is little consensus in relative rankings between the three ranking strategies which is somewhat surprising in the case of confidence versus index rankings given that examinee confidence is incorporated in the calculation of the targeted index. Pairwise correlations are not strong especially between funnel rankings and the other two. The median difference between funnel and confidence rankings and funnel and index rankings is -1 and 3, respectively. This means that topics ranked by index tend to be higher than when ranked by funnel whereas funnel and confidence rankings are

¹We exclude Emergent Topics (Legal, Ethical, DEI) at 2%.

Medical content category	Allocation	Topic	Funnel rank	Mean difficulty
Cardiovascular System	12%	Aortic aneurysm, abdominal	1	1.262
Dermatologic System	5%	Erythema multiforme	24	1.327
Endocrine System	8%	Type 2 diabetes mellitus/hyperosmolar hyperglycemic syndrome	4	2.765
Eyes, Ears, Nose & Throat	8%	Eustachian tube dysfunction	7	0.982
Gastrointestinal System/Nutrition	10%	Constipation	13	1.695
Genitourinary System	5%	Benign prostatic hyperplasia	3	0.895
Hematologic System	4%	Hypocoagulable states	5	0.495
Infectious Diseases	7%	Septic disorders: sepsis, septic shock, systemic inflammatory response syndrome	13	1.871
Musculoskeletal System	8%	Osteoarthritis	13	1.612
Neurologic System	5%	Major neurocognitive disorder/dementia	7	0.732
Psychiatry/Behavioral Science	7%	Disruptive, impulse-control, and conduct disorders: oppositional defiant disorder, conduct disorder	20	1.379
Pulmonary System	10%	Asthma	18	2.038
Renal System	4%	Renal cell carcinoma	19	1.558
Reproductive System	5%	Pregnancy loss	16	0.594

Table 1: List of the top ranked topics per organ system including the mean topic difficulty and each organ system's percent allocation on the blueprint. We highlight the **easiest** and **most difficult** topic by mean difficulty; and the **highest rank** and **lowest rank** per topic.

Medical content category	Topic	Funnel rank	Index rank	Confidence rank	Mean difficulty
Cardiovascular System	Aortic aneurysm, abdominal	1	28	82	1.262
Cardiovascular System	Venous insufficiency	1	25	24	1.450
Infectious Diseases	Rubella	34	1	41	2.991
Endocrine System	Thyroiditis	34	19	1	2.082
Neurologic System	Bell palsy	233	120	114	-0.122
Infectious Diseases	Rabies	259	237	237	-3.238

Table 2: The top portion lists the top ranked by each ranking system and the bottom portion the lowest ranked. Note that rabies is the lowest ranked for both confidence and index rankings.

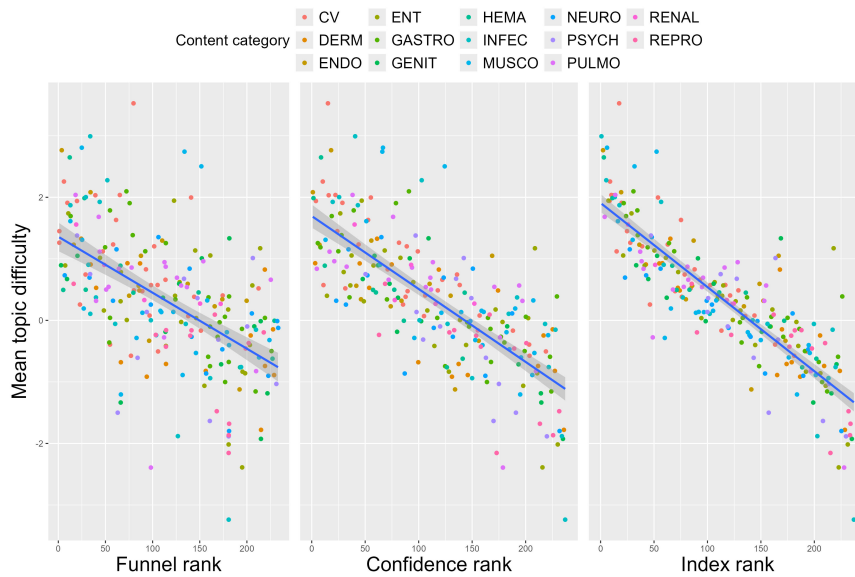


Figure 2: A plot of the correlation of funnel rankings, confidence rankings and index rankings against mean topic difficulty. We also include regression lines over all topics including confidence intervals.

closer on average. Combined with the weak correlation, the different ranking types have some similarity in ordering though there are some topics with large differences in ranking. A natural question arises about how the combined information from the rankings informs us about the behavior of examinees. For example, for a topic that is ranked low by confidence and high by funnel might indicate an area where examinees tend to be overconfident or a similar ranking between confidence and index could indicate self-reported confidence better aligned with performance.

An implication related to this is uncovering different subpopulations within the examinees by their performance in relation to the topic rankings. Currently we do not distinguish responses between different groups of examinees. Since all examinees are given targeted items even if their prior quarter was perfect, this creates the concern that performance on targeted items may be inflated due to the contribution of responses of high performing examinees. Topics with similar rankings across the ranking types might indicate more true knowledge gaps relative to the examinee population. On the other hand, topics with more discordant rankings might suggest weaker or stronger knowledge gaps among different subgroups. One possibility would be to create new rankings according to these separate groups which could be useful to have a more accurate understanding of how groups of examinees are performing. In an operational

application, this could lead to different ways of prioritizing topics for emphasis across examinees.

Since each type of ranking has modest at best correlation with item difficulty, further analysis of the items themselves could shed light on the rationale. Traditional NLP methods like feature engineering on textual attributes have shown middling success at categorizing item difficulty (Contreras, 2018; Ha et al., 2019) and even language models fine-tuned for the biomedical domain have difficulty in improving on difficulty prediction compared to general domain models (Li et al., 2025). This suggests there may be a more metacognitive rationale explaining these performance gaps that are not captured in the text itself. Additionally we only analyze rankings at the topic level, separating into topic-task combinations might provide more information that is lost averaging over topics.

In summary, the results of this study provide a new technique for relative comparison of performance across topics within a longitudinal examination. Practical applications such as incorporation into recommender systems or personalized study plans over emphasized topics could lead to overall increased performance or earlier exits from the LA.

Limitations

A major limitation to the current study is the structure of the data. Although there is a healthy

Medical content category	Topic	Funnel rank	Index rank	Confidence rank	Mean difficulty
Cardiovascular System	Venous insufficiency	1	25	24	1.450
Cardiovascular System	Aortic aneurysm, abdominal	1	28	82	1.262
Genitourinary System	Benign prostatic hyperplasia	3	33	62	0.895
Endocrine System	Type 2 diabetes mellitus/hyperosmolar hyperglycemic syndrome	4	2	18	2.765
Hematologic System	Hypocoagulable states	5	71	57	0.495
Cardiovascular System	Coronary artery disease	6	9	10	2.256
Neurologic System	Major neurocognitive disorder/dementia	7	43	52	0.732
Eyes, Ears, Nose & Throat	Eustachian tube dysfunction	7	41	58	0.892
Genitourinary System	Overactive bladder	9	45	48	0.672
Cardiovascular System	Myocardial infarction, acute: non-ST-segment elevation myocardial infarction (NSTEMI)	10	20	29	1.910

Table 3: Top 10 topics overall by funnel rank.

volume of responses across examinees, the longitudinal range is only three years. This limits our comparisons across administration windows and our current trends might not be stable across a longer timescale.

Another limitation is that confidence and relevance scores are self-reported. Since examinees interact with the scale individually reported scores are not necessarily equivalent across examinees. This has the possible impact of skewing the interpretation of the scores where, for example, a topic is judged to have higher average examinee confidence while having relatively low performance. For reliability it would be helpful to equate them on the same scale.

References

- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Luis A. Contreras. 2018. An application of natural language processing for triangulation of cognitive load assessments in third level education. Master’s thesis, Technological University of Dublin.
- Scott H Fraundorf, Zachary A Caddick, Timothy J Nokes-Malach, and Benjamin M Rottman. 2023. Cognitive perspectives on maintaining physicians’ medical expertise: IV. best practices and open questions in using testing to enhance learning and retention. *Cogn. Res. Princ. Implic.*, 8(1):53.
- Sarah E Giron, Deniz Dishman, Susan P McMullan, Jared Riel, Timothy Newcomer, Dennis Spence, and Shahid A Choudhry. 2021. Longitudinal assessment: A strategy to improve continuing professional certification. *J. Prof. Nurs.*, 37(6):1140–1148.
- Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. [Predicting the difficulty of multiple choice questions in a high-stakes medical exam](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20, Florence, Italy. Association for Computational Linguistics.
- Marian Hickendorff, Willem J Heiser, Cornelis M van Putten, and Norman D Verhelst. 2009. How to measure and explain achievement change in large-scale assessments: A rejoinder. *Psychometrika*, 74(2):367–374.

- Enshuo Hsu and Kirk Roberts. 2025. LLM-IE: a python package for biomedical generative information extraction with large language models. *JAMIA Open*, 8(2):ooaf012.
- Sunbeom Kwon and Susu Zhang. 2025. Explaining performance gaps with problem-solving process data via latent class mediation analysis. *Psychometrika*, 90(5):1–29.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Ming Li, Hong Jiao, Tianyi Zhou, Nan Zhang, Sydney Peters, and Robert W Lissitz. 2025. Item difficulty modeling using fine-tuned small and large language models. *Educ. Psychol. Meas.*, 85(6):1065–1090.
- Ying Liu and Steven Culpepper. 2025. Designing learning intervention studies: Identifiability of heterogeneous hidden markov models. *Psychometrika*, 90(4):1258–1283.
- Bengt O Muthén and Siek-Toon Khoo. 1998. Longitudinal studies of achievement growth using latent variable modeling. *Learn. Individ. Differ.*, 10(2):73–101.
- Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.*, 4(1):86.
- Henry L Roediger and Jeffrey D Karpicke. 2006. Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.*, 17(3):249–255.
- Daniel P Schmidt. 2020. Identifying knowledge gaps using a graph-based knowledge representation. Master’s thesis, Wright State University.
- Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Pérez Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024. [The BEA 2024 shared task on the multilingual lexical simplification pipeline](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.
- David Spiegelhalter, Christopher Sherlaw-Johnson, Martin Bardsley, Ian Blunt, Christopher Wood, and Olivia Grigg. 2012. Statistical methods for healthcare regulation: Rating, screening and surveillance. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 175(1):1–47.
- David J Spiegelhalter. 2005. Funnel plots for comparing institutional performance. *Stat. Med.*, 24(8):1185–1202.
- Matthias von Davier, Xueli Xu, and Claus H Carstensen. 2011. Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, 76(2):318–336.
- Robyn C Ward, Kathy A Baker, Dennis Spence, Colleen Leonard, Alysha Sapp, and Shahid A Choudhry. 2023. Longitudinal assessment to evaluate continued certification and lifelong learning in healthcare professionals: A scoping review. *Eval. Health Prof.*, 46(3):199–212.
- Isabella Catharina Wiest, Fabian Wolf, Marie-Elisabeth Leßmann, Marko van Treeck, Dyke Ferber, Jiefu Zhu, Heiko Boehme, Keno K Bressemer, Hannes Ulrich, Matthias P Ebert, and Jakob Nikolas Kather. 2025. A software pipeline for medical information extraction with large language models, open source and suitable for oncology. *NPJ Precis. Oncol.*, 9(1):313.
- Kazuhiro Yamaguchi, Haruhiko Mitsunaga, Shun Saso, and Yuri Uesaka. 2026. Predicting multilevel growth trajectories using a random-effect diagnostic classification model. *Psychometrika*, pages 1–46.