



The Score

October 2026
Volume 48, Issue 4

Editor:
Kevin M. Williams

APA Division 5 - Quantitative and Qualitative Methods

INSIDE

Feature Article

Less is More? Evaluating Shorter Versions of an Open-Response Situational Judgement Test
(Robb & Sitarenios, p. 1)

In This Issue

Message from the APA Division 5 President (Wilson, p. 10)
A Review of Division 5 Programming at the 2026 APA Convention (Pek et al., p. 11)
Congratulations to the Latest APA Division 5 Fellows! (p. 14)
Publication Corner (p. 14)
Links to Previous Issues (p. 16)
Div. 5 Membership Info (p. 17)
Advertising in The Score (p. 17)
Publishing in The Score (p. 17)



Less is More? Evaluating Shorter Versions of an Open-Response Situational Judgement Test

Colleen Robb & Gill Sitarenios

Acuity Insights, Toronto, Ontario, Canada

crobb@acuityinsights.com

A case study identifying optimal test length via question-level scoring

1. Introduction

Test length reduction is not a new concept or rare phenomenon. Since shortened tests require less time and resources (Stanton et al., 2002) and have the propensity to reduce respondent burden (Smits & Vorst, 2007), it's no surprise that test developers often look for ways to reduce test length. However, these potential benefits also bring potential risks, namely, the possibility of reduced psychometric quality of the assessment. These

concerns regarding psychometric quality are not without merit. A literature review of 164 shortened tests indicated that the mean reliability of tests reduced from $\alpha=0.84$ to $\alpha=0.77$ when test lengths were reduced and that validity was only assessed in approximately half of the sample (Kruyen et al., 2013). Further, much of the conversation around the shortening of tests or assessments is tied to multiple choice or forced choice formats. There is virtually no literature available on how one may reduce the length of an open-response test.

The goal of this study was to explore optimal test length, which balances practical utility with psychometric integrity (Sitarenios, 2022), for an open-response situational judgement test (SJT).

2. Study Design

To explore this research question, we employed a two-phase sequential design to examine the Casper test; a 14-item open-response SJT that contains both typed-response and video-response scenarios. Each of the 14 items were unique scenarios that each asked two unique questions. Given that this instrument was relatively short already, we identified an opportunity to also explore the impact of changing from scenario-level scoring to question-level scoring. We hypothesized that changing from scenario-level scoring to question-level scoring would have two main advantages: (1) test takers would receive more accurate scores since they were being evaluated for each individual response opposed to holistically for all questions in a scenario and (2) it would allow more flexibility in plausible test length options given that we'd essentially be doubling the number of items in the test.

The study was organized into two main phases. In the first phase, trained raters were asked to provide scenario-level scores and question-level scores for a set of data so that we could assess the extent to which scores differed between the two methods. Using the results of this phase, we conducted a series of simulations across a larger dataset to evaluate the psychometric properties of the test with question-level scoring across various test lengths.

2.1 Phase 1

2.1.1 Procedure

We employed a repeated-measures design in which participants scored the same response twice across the two methods: scenario-level and question-level. The order in which the methods were presented to participants were randomized to control for order effects. Participants were not permitted to score the responses a second time (the alternate approach) until they participated in scoring an active test (i.e., a real test outside the study environment). Scoring an additional set of unique responses between study sessions was intended to reduce the likelihood that participants would recall and reuse scores from the first round when scoring the responses a second time. Response order was also randomized between rounds to further limit recall of prior scores.



Colleen Robb, MA
Acuity Insights



Gill Sitarenios, PhD
Acuity Insights

2.1.2 Sample

We wanted to ensure that we captured data from a diverse sample of tests since the Casper test is used globally, in different languages, and for admissions to a variety of professional programs. Ultimately, we employed $n=23$ trained raters to examine data from $n=200$ test takers across the United States, Australia, and Canada (English and French). Details on the sample data is available in Table 1.

Table 1. Information Regarding Sample Data in Phase 1

Country	Language	Program Type	Number of Raters	Number of Test Takers
United States	English	Health Sciences	8	50
Canada	French	Allied Health	4	50
Australia	English	Teacher's Education	6	50
Canada	English	Health Sciences	5	50

2.1.3 Results

Descriptive Statistics of Total Scores. Descriptive statistics for both scenario-level and question-level scores are provided in Table 2.

Table 2. Descriptive Statistics of Total Scores Across Scoring Methods

	Scoring Method	<i>n</i>	mean	<i>sd</i>	median	min	max	Skew	Kurt
US	Scenario-Level	50	5.47	0.94	5.62	2.75	7.00	-0.79	0.44
US	Question-Level	50	5.31	0.90	5.41	2.69	6.88	-0.49	-0.06
AUS	Scenario-Level	50	4.49	0.95	4.50	2.17	7.00	-0.08	-0.02
AUS	Question-Level	50	4.12	0.93	4.08	1.92	6.25	0.14	-0.30
CA-E	Scenario-Level	50	5.01	1.12	5.00	2.40	7.20	-0.01	-0.64
CA-E	Question-Level	50	5.02	0.96	5.00	3.00	7.20	0.26	-0.39
CA-F	Scenario-Level	49	5.96	1.03	6.00	3.50	7.75	-0.34	-0.45
CA-F	Question-Level	49	5.42	1.09	5.50	2.62	7.75	-0.30	-0.14

Note. One CA-F test taker had to be removed as a test taker did not complete the assessment.

Absolute Mean Difference in Total Scores Across Scoring Methods. We examined the difference in total scores between the two methods. To do this, we calculated the mean absolute difference between test takers' total scores from the scenario-level scoring and their total scores from the question-level scoring. Descriptive statistics are available in Table 3 and a frequency table of absolute differences is available in Table 4.

Table 3. Descriptive Statistics of Mean Absolute Differences in Total Scores

Country	<i>n</i>	mean	<i>sd</i>	median	min	max	Skew	Kurt
US	50	0.36	0.24	0.31	0.00	0.88	0.40	-0.89
AUS	50	0.50	0.38	0.46	0.00	1.42	0.48	-0.58
CA-E	50	0.45	0.38	0.40	0.00	1.60	0.88	0.40
CA-F	49	0.64	0.46	0.50	0.00	1.75	0.49	-0.86

Relationship Between Total Scores Across Scoring Methods. The correlation evidenced a strong and statistically significant relationship between the two total scores in the US data ($r=0.87$, $p<.001$), Australian data ($r=0.86$, $p<.001$), Canadian English data ($r=0.85$, $p<.001$), and Canadian French data ($r=0.82$, $p<.001$).

2.1.4 Phase 1 Discussion

Across all data, we found that test takers' total scores differed, on average, by less than 1 point (range: 0.36 - 0.64) between the two methods. More specifically, 60% of test takers' total scores differed by half a point or less and only 10% of test takers' total scores differed by more than one point. Correlation analyses evidenced a strong relationship between the two scores ($r=0.82$ - 0.87 , $p<.001$). Together, these results helped inform the second phase of the study.

2.2 Phase 2

2.2.1 Procedure

In the second phase, we employed a synthetic design approach similar to what is described by Lee and colleagues (2014). Scenario-level scores were collected from the full-length version of the open-response SJT and then used to generate synthetic question-level scores. From here, we were able to evaluate the psychometric properties of various test lengths without the practical and ethical disadvantages of asking participants to take the test repeatedly across different lengths as would be required with a traditional experimental design (Lee et al., 2014). To generate the synthetic question-level scores, we wrote a script using RStudio Version 2023.3.0.386 (Posit Team, 2023), which generated two item-level scores for each scenario that differed by +/- one point from their original scenario-level score. The choice to allow scores to differ within one point of the original score was determined based on findings in Phase 1.

Table 4. Frequency Table of Absolute Differences in Total Scores

	US		AUS		CA-E		CA-F	
Absolute Difference	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
0 - 0.10	10	20	11	22	12	24	2	4.08
0.11 - 0.20	4	8	3	6	7	14	7	14.29
0.21 - 0.30	7	14	2	4	4	8	7	14.29
0.31 - 0.40	11	22	4	8	6	12	5	10.20
0.41 - 0.50	5	10	7	14	2	4	4	8.16
0.51 - 0.60	3	6	6	12	6	12	0	0
0.61 - 0.70	5	10	3	6	2	4	3	6.12
0.71 - 0.80	2	4	3	6	4	8	3	6.12
0.81 - 0.90	3	6	2	4	2	4	4	8.16
0.91 - 1.00	0	0	4	8	2	4	3	6.12
1.01 - 1.10	0	0	3	6	1	2	0	0
1.11 - 1.20	0	0	0	0	0	0	3	6.12
1.21 - 1.30	0	0	0	0	0	0	5	10.20
1.31 - 1.40	0	0	0	0	1	2	1	2.04
1.41 - 1.50	0	0	2	4	0	0	0	0
1.51 - 1.60	0	0	0	0	1	2	0	0
1.61 - 1.70	0	0	0	0	0	0	1	2.04
1.71 - 7.80	0	0	0	0	0	0	1	2.04

Prior to conducting the simulations, we identified a number of different test length combinations to explore. Although there were 42 possible combinations in theory, several were not practically feasible. Therefore, we removed several combinations prior to conducting our analyses. Ultimately, we examined 15 possible test length combinations (labeled with letters in Table 5). Data analyses were conducted on each test length combination across each independent test instance.

Table 5. Possible Test Length Combinations

Number of Typed Responses	Number of Video Responses					
	1	2	3	4	5	6
	1					
	2					
	3			A	B	C
	4			D	E	F
	5			G	H	I
	6			J	K	L
	7			M	N	O
	8					

2.2.2 Sample

This phase utilized data from 26,671 test takers from 20 unique test instances across the United States, Australia, and Canada (English and French). See Table 6 for details on the data used for Phase 2.

Table 6. Information Regarding Sample Data in Phase 2

Country	Language	Number of Tests Examined	Number of Test Takers
United States	English	7	10,545
Australia	English	5	5,036
Canada	English	5	7,572
Canada	French	3	3,518

2.2.3 Results

For each combination across each test instance, we examined test-level reliability, the correlation between the total score and the original total score from the 14-item test, and demographic group differences.

Internal Consistency Reliability. Average reliability for the original 14-item test was $\alpha=0.82$ with an average inter-item correlation (IIC) value of 0.24. Average reliability and IIC values for each combination across all test instances are available in Table 7.

Table 7. Average Alpha and IIC Values Across Test Instances

Test Length Combination	Average Alpha	Average IIC
Original	0.82	0.24
A	0.77	0.22
B	0.77	0.20
C	0.78	0.18
D	0.80	0.23
E	0.80	0.20
F	0.81	0.19
G	0.83	0.23
H	0.83	0.21
I	0.83	0.20
J	0.85	0.24
K	0.85	0.22
L	0.85	0.21
M	0.87	0.25
N	0.87	0.23
O	0.87	0.22

Correlations. Correlation coefficients were used to provide an estimate of the strength of the relationship between the original total score and the total score for each combination. The average correlation coefficient for each test length combination is available in Table 8.

Table 8. Average Correlation Between Original Scores and Various Test Length Combinations Across Test Instances

Test Length Combination	Average Correlation
A	0.85
B	0.87
C	0.88
D	0.89
E	0.90
F	0.91
G	0.91
H	0.92
I	0.93
J	0.93
K	0.94
L	0.95
M	0.94
N	0.95
O	0.96

Demographic Group Differences. Demographic differences were explored in aggregate for each grouping (US, Australia, Canada English, and Canada French) to ensure a sufficient sample size. We calculated the magnitude of group differences for the total scores of the original 14-item test and compared them to the magnitude observed in the total scores from each test length combination. Generally, we observed very minimal variability in the magnitude of group differences between the original score and scores from the various test lengths. More specifically, the difference in magnitude was no greater than 0.06 across all group comparisons in all geographies.

2.2.4 Phase 2 Discussion

Overall, a majority of the test length combinations performed well. Of the 15 combinations explored, 60% (n=9) produced an average reliability that exceeded that of the original test format. Additionally, 73.3% (n=11) met or exceeded the correlation coefficient threshold of 0.90 that suggests the shortened test adequately represents its full length counterpart (Lee et al., 2014).

3. Discussion

This study explored whether an open-response situational judgment test (SJT) could be shortened while maintaining psychometric quality. Across both phases, the findings were encouraging. Phase 1 showed that question-level scoring produced scores very similar to scenario-level scoring with average correlations of $r=0.82$ to 0.87 ($p<.001$) between the two methods. At the test taker level, we saw that their total scores differed, on average, by less than 1 point between the two methods. This finding suggests that scoring each question individually does not dramatically change the overall results, but it does provide more flexibility in structuring the test. Phase 2 expanded on the Phase 1 results by simulating various test lengths using synthetic question-level data. The majority of shortened versions maintained reliability greater than what was observed in the original 14-item test. Further, correlations with the original test were high, indicating that the shortened tests adequately represented the longer version. Also, the change in demographic group differences were minimal, suggesting that shortening the test did not introduce bias across groups.

Taken together, these findings have practical implications. Shortening an open-response SJT can reduce test-taking time and respondent burden without sacrificing psychometric quality. Additionally, question-level scoring allows more granular measurement, enabling test developers to explore a wider range of test lengths while maintaining psychometric standards. Of course, there are limitations. Our simulations relied on synthetic data, which, while based on real scoring patterns, may not fully capture all nuances of real-world responses. Further, this study examined only a single instrument. It is recommended that future studies examine different instruments with different formats.

Ultimately, the study demonstrates that open-response tests can be made more efficient without a major trade-off in reliability or validity. This opens the door for more practical and user-friendly assessments in higher-education admissions and other high-stakes testing environments.

4. References

- Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, 13(3), 223-248. <https://doi.org/10.1080/15305058.2012.703734>
- Lee, E. S., Bygrave, C., Mahar, J., & Garg, N. (2014). Can higher education exams be shortened? A proposed methodology. *International Journal of Information and Education Technology*, 4(6), 517-525. <https://doi.org/10.7763/IJiet.2014.V4.462>
- Posit Team. (2023). *RStudio* (Version 2023.3.0.386) [Computer software]. Posit Software, PBC. <https://posit.co/>
- Sitarenios, G. (2022). Short versions of tests: Best practices and potential pitfalls. *Journal of Pediatric Neuropsychology*, 8(3), 101-115.
- Smits, N., & Vorst, H. C. M. (2007). Reducing the length of questionnaires through structurally incomplete designs: An illustration. *Learning and Individual Differences*, 17, 25-34. <https://doi.org/10.1016/j.lindif.2006.12.005>
- Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, 55, 167-194. <https://doi.org/10.1111/j.1744-6570.2002.tb00108.x>

Message from the APA Division 5 President

Leo Wilton, Ph.D.

Binghamton University

lwilton@binghamton.edu



It is truly an honor to serve as the APA Division 5 president for the 2026-2027 term. Together, we can foster collaboration, innovation, and a strong sense of community within our division. I look forward to working alongside the Division 5 Executive Committee, Sections, Committees, membership, and our broader community. I am dedicated to bringing together our methodological perspectives in EMS, SQIP, and assessment in building a strong community in our division.

Building on the work of many pioneers in Division 5, I will bring 20+ years of demonstrated leadership, scientific research, education, and policy expertise to support the Division's growth and impact in psychological science and research methods. My Division interests include (1) fostering collaborative interdisciplinary research in a social justice context, (2) enhancing membership growth and visibility within the U.S. and globally and building relationships with APA divisions, (3) strengthening programmatic research efforts (e.g., symposiums), (4) integrating the voices of the younger generation into the Division's leadership structure through mentoring collaborations, and (5) strengthening communication of research methodologies to the general public.

I have exciting news to share in that we have the forthcoming *Handbook of Methodology for Psychology in the 21st Century: Progress, Critique, and Future* to be published by Guilford Press. This groundbreaking project is being led by Fred Wertz, Lisa Harlow, and Rachel Fouladi alongside a talent team of section editors. The comprehensive volume will span a wide range of methodological domains, including Qualitative Inquiry, Quantitative Analysis, Measurement and Assessment, Evaluation, and Mixed methods. We have some programmatic efforts related to the handbook underway. Stay tuned for updates.

We have had a strong Division 5 presence in APA initiatives, including the APA Town Hall on Artificial Intelligence (AI) that featured presentations by Dr. Arthur Evans, Dr. Tony Habash, Dr. Vale Wright, and Corbin Evans on APA's approach to AI in three areas, including internal operations, psychological field applications, and public policy. APA's strategic priorities focus on using AI as a tool to augment rather than replace psychological work, addressing potential harms while maximizing benefits, and establishing a virtual center for behavioral science and AI to fill the missing "human layer" in AI development. The discussion covered various AI applications in healthcare, workplace impacts, regulatory efforts, and environmental considerations, with participants raising questions about data center environmental impacts and the balance between profit motives and psychological science implementation in AI companies.

A Review of Division 5 Programming at the 2026 APA Convention

Jolynn Pek, Ph.D.¹, Ken Kelley, Ph.D.², & Zack Fisher, Ph.D.³

¹The Ohio State University, ²University of Notre Dame, ³The University of North Carolina at Chapel Hill

pek.5@osu.edu

Psychological Methods with Artificial Intelligence

Psychological methods is a broad and pluralistic discipline that encompasses the philosophy of science, qualitative inquiry, statistics, measurement, assessment, modeling, and program evaluation. Over the past year, Division 5 initiatives, events, and programs have sought to advance methodology rather than any particular *method*. Stated differently, our efforts have emphasized the principles, values, and reasoning that guide the use of our scientific tools across diverse contexts. This perspective builds on the vision advanced by past Division 5 Presidents [Fred Wertz](#) and [Alex Beaujean](#), who highlighted the importance of methodological reflection, intellectual openness, and respect for diverse approaches as catalysts for innovation and the generation of new knowledge in psychological science and practice.

Recognizing the transformative nature of Artificial Intelligence (AI) on research and practice, 2026 APA Division 5 Program Chairs **Ken Kelley** and **Zack Fisher** organized five timely, informative, and well-attended convention sessions, each offering continuing education credit. These sessions addressed critical questions facing the field today: How is AI affecting data quality? In what ways can AI enhance assessment processes? How can AI support scientific discovery and qualitative inquiry? What might be potential pitfalls? And more broadly, how are psychological methods evolving in response to advances in AI? Together, these discussions highlighted both the opportunities and challenges that AI presents for methodological research and practice.

The five sessions were:

I. AI, Survey Responding, and Data Quality

This session examined how advances in AI are reshaping survey research and raising new questions about data quality. **Andrew Gordon** (Prolific) reported proprietary research platforms (e.g., Prolific, CloudResearch Connect, and Verasight) generally yield higher quality data than hybrid panels (e.g., Prodege, Dynata, Qualtrics) or aggregator panels (e.g., Clint, PureSpectrum, Primepanels). Although fully autonomous AI agents can generate survey responses, their large-scale use remains limited by cost and technical barriers. More concerning is the rise of AI-assisted human responding, as low-quality human data currently presents a greater threat to research validity than fully AI-generated responses. **Nick Stagnaro** (MIT), reached similar conclusions based on his work developing tools to detect AI-generated survey answers. **Marsha Krupenkin** (University of



Jolynn Pek, PhD
The Ohio State University



Ken Kelley, PhD
University of Notre Dame



Zack Fisher, PhD
The University of North Carolina
at Chapel Hill

Maryland) demonstrated that features of open pen-ended survey responses, such as device type and response mode (voice vs. text), can influence the quality of large language model (LLM) analyses. These differences can affect substantive conclusions and might be amplified by automated text-analysis procedures.

II. Using generative AI to create or analyze survey or test items

This session examined the growing role of generative AI in assessment. **Kevin Williams** (ETS) discussed how AI is being integrated into assessment workflows, including content generation, item bank maintenance, automated scoring, and simulation-based assessment and training. Although **Alison Cheng** (University of Notre Dame) was unable to present in person, her [materials](#) provided a historical overview of generative AI (GenAI) in assessment, highlighting its roles as an item developer, respondent, judge/rater, and interactive agent.

Both presentations emphasized that GenAI can support assessment development, training, and evaluation, but increased model sophistication alone (e.g., use of LLMs in place of traditional psychometric models) does not guarantee better outcomes. Data quality remains paramount. A central theme was that AI should augment, rather than replace human expertise. Human oversight is essential to ensure validity, transparency, fairness, interpretability, privacy, security, and meaningful variability in assessment. As discussant, **Ken Kelley** (University of Notre Dame) noted that AI has long played a role in assessment, particularly through computerized adaptive testing, and that recent advances hold considerable promise for further strengthening psychological science and practice.

III. Harnessing AI for Scientific Discovery: Quantitative and Qualitative Perspectives

This session explored how AI is expanding opportunities for scientific discovery across both qualitative and quantitative research. **Gabriel Velez** (Marquette University) discussed the growing use of GenAI for coding, memoing, summarizing, and interpretation in qualitative research. Drawing on a relational perspective that views GenAI as a third presence between researcher and participant, he highlighted risks such as cognitive offloading, premature interpretive closure, threats to confidentiality, and erosion of community trust. To address these concerns, he advocated for human-led approaches, routine checks for bias and hallucinations, and clear standards for data governance.

From a quantitative perspective, **Peter Kvam** (The Ohio State University) focused on the scientific potential of deep learning beyond LLMs. He argued that machine learning (ML) methods excel at prediction, classification, and dimension reduction, making them particularly useful for estimating and comparing highly complex models. Grounded in the universal approximation theorem, neural networks can approximate any smooth relation between predictors and outcomes when sufficiently parameterized. As a result, ML methods offer powerful tools for uncovering novel patterns and representations in data, facilitating theory development and scientific discovery.

IV. Psychological Methodology Meets Artificial Intelligence: Modern Innovations and Future Advancements

The keynote session highlighted the growing convergence of psychological methodology and AI. **Ahmed Abbasi** (University of Notre Dame) described three AI-enabled paradigms relevant to psychological science: (a) using pretrained models to infer psychological constructs from text, (b) prompting GenAI models to adopt specific roles and perspectives, and (c) evaluating AI performance using psychometric methods. He emphasized that realizing the full potential of these approaches will require closer collaboration between psychological methods and computational scientists.

Kai Larsen (University of Colorado – Boulder) demonstrated how AI can translate theory into practice through [PsyProxy](#), a platform that analyzes text to identify theoretically grounded constructs linked to outcomes of

interest. Drawing on frameworks from psychology and related social and behavioral sciences, the platform can be applied to diverse text sources, including interviews, reviews, patient narratives, and AI-generated conversations. Such tools can help uncover hidden patterns in text, support business and health analytics, and distinguish between human and AI-generated content.

V. Presidential Address: Reflections on Psychological Methodology and AI

Drawing together themes from across the program, **Jolynn Pék** (The Ohio State University) reflected on the intertwined histories of psychological methodology and AI. She noted that modern AI emerged from advances in statistics, ML, Bayesian computation, and data science, while psychology has long contributed to AI through research on learning, cognition, and human behavior. Learning theory helped shape ML, neural networks were inspired by biological neural systems, and cognitive science – an interdisciplinary field integrating psychology, computer science, linguistics, neuroscience, and philosophy – played a central role in shaping modern AI.

Pék observed that AI is transforming every stage of the scientific process. Viewed through the distinction between *method* and *methodology*, ML approaches have largely been treated as a method for data analysis. GenAI, however, increasingly functions as an active collaborator in research workflows, contributing to construct development, measurement, data generation, coding, interpretation, and evaluation, as reflected throughout the convention program. This shift raises important methodological questions about how human-AI partnerships should be structured and governed in psychological science and practice.

Looking ahead, AI presents new opportunities to advance longstanding methodological debates, including prediction versus explanation, nomothetic versus idiographic approaches, and the relation between observed data and latent psychological constructs. At the same time, AI raises fundamentally psychological questions about agency, identity, cognition, and lived experience in human-AI interactions. For psychological methodologists and scientists, these developments represent not only a technological frontier but also a rich agenda for future inquiry.

Congratulations to the Latest APA Division 5 Fellows!

Division 5 is proud to welcome the following individuals into Fellowship:

Winfred Arthur, Jr., PhD, holds the John Paul Abbott Professorship in Liberal Arts and is a professor in the Department of Psychological and Brain Sciences at Texas A&M University. He has published and consulted extensively in the areas of testing, assessment, personnel selection, validation, individual differences, training, and research methods. He is a Fellow of the Society for Industrial and Organizational Psychology, Association for Psychological Science, and American Psychological Association.

Jen Katz-Buonincontro, PhD, serves as Professor, Policy, Organization and Leadership and Courtesy Professor, Psychological and Brain Sciences at Drexel University. Her passion is helping people unlock their creative potential. She teaches, mentors and supervises in three areas- research methods, creativity and leadership development. A current example NSF research project focuses on AI for creative problem solving in teacher and student communication. Dr. Katz-Buonincontro also serves as Vice program chair, Human research committee with the Ethics division (APA). She loves combining research methods in new ways to capture fluid, changing and complex interactions, both human and technological.

Eun-Young (E.Y.) Mun, PhD, CPH, is a Regents Professor and Associate Dean for Research at the University of North Texas Health at Fort Worth. Since 2010, Dr. Mun has directed the NIH-funded *Project INTEGRATE* and recently launched the *Project INTEGRATE Data Science Academy* to train emerging scientists. Her research strengthens the evidence base for alcohol prevention and intervention through rigorous data synthesis and advanced analytics. A champion of open science and collaborative team science, she leads interdisciplinary efforts to improve critical health outcomes and advance quantitative methods while mentoring the next generation of scientists for an artificial intelligence enabled future.

Publication Corner



Freuli, F., Noventa, S., & Lombardi, L. (2026). New frontiers in the p-hacking practices: Assessing the effect of data peeking practice on type-I error rate through computational modeling. *Frontiers in Psychology*, 17, 1769309. <https://doi.org/10.3389/fpsyg.2026.1769309>

Data-peeking, the practice of stopping data collection once significant results are obtained, poses a significant threat to the credibility of research by increasing the likelihood of observing and publishing false positive results. The effect of this practice on scientific reliability has traditionally been assessed through Monte-Carlo simulations, which suffer from some methodological issues (e.g., Monte-Carlo standard error is often overestimated). This paper presents a computational model that, through a series of convolutions between normal and truncated normal distributions, estimates the increase in false-positive rates due to data-peeking practice, overcoming the issues of the simulation method. The performance of the computational model was compared against Monte-Carlo simulations under different conditions, including alpha levels (0.05, 0.025, 0.01, and 0.001), sample sizes (40, 80, 160, and 320), and numbers of interim analyses (1 to 10). The results confirmed the computational model's accuracy and reliability, showing that computational values fell within two standard deviations of the simulated means. In conclusion, the proposed model not only presents a faster and error-reducing approach to estimating the effects of data-peeking practice but also provides guidelines for the

development of similar models to examine other forms of p-hacking, thereby improving the understanding and study of their effects.

Krumm, S., Thiel, A. M., Reznik, N., Freudenstein, J.-P., Schäpers, P., & Mussel, P. (2026). Creating a psychological test in a few seconds: Can ChatGPT develop a psychometrically sound situational judgment test? *European Journal of Psychological Assessment*, 42(4), 344-351. <https://doi.org/10.1027/1015-5759/a000878>

The development of Situational Judgment Tests (SJTs) is a multi-step procedure that often requires substantial resources and time. ChatGPT as a large language model has already proven helpful for many complex tasks, including the development of psychological questionnaires. However, ChatGPT has so far not been tested for its ability to develop SJT items, which could speed up SJT development and thus contribute to the dissemination of psychometrically well-designed SJTs in practice and also support research requiring specific SJTs or SJT versions. Thus, the current study ($N = 419$) examined whether ChatGPT (2023, Feb 13 version using ChatGPT-3.5) can create SJT items, assessing a facet of personality (i.e., gregariousness), that show similar psychometric properties (examined through confirmatory factor, reliability, and correlational analyses) as human-created SJT items. Results revealed that the measurement model for all eleven SJT items that were created by ChatGPT showed a poor model fit, while a reduced set of eight items yielded an acceptable fit. Reliability estimates as well as convergent and discriminant validity evidence were similar for the human-created and the ChatGPT-created SJT. Implications as well as potential boundary conditions of SJT development through ChatGPT are discussed.

Østergaard Jaspers, S., Weinke Christiansen, S., Dybdal Angersen, E., Martin, A.-M., & Madsen, I. E. H. (2026). Collective capacity building for emotionally charged work: A grounded theory of strategies for maintaining meaning and well-being in emotionally charged work. *International Journal of Qualitative Studies on Health and Well-Being*, 21(1). <https://doi.org/10.1080/17482631.2026.2722429>

Emotionally charged work – encompassing emotional demands and emotional labour – has been associated with ill health and alienation. While the consequences of such work has been widely theorized, the prevention of negative consequences remains undertheorized. In this study, we adopt a classical grounded theory approach to (a) Explore the main concern of employed people regarding emotionally charged work and (b) Generate a grounded theory that could inform sustainable organization of emotionally charged work and the prevention of compassion fatigue. Initial purposeful sampling included eight employees and managers from psychiatric hospitals. As subsequent theoretical sampling required broader diversity, we conducted interviews with 15 additional participants across varied occupations, including ministers of religion, military personnel, social care workers, educators, IT managers, and others. Using open coding and constant comparison, we developed a grounded theory of collective capacity building for emotionally charged work. We found that the main concern of participants was to maintain meaningful working lives and in the context of emotionally charged work, they achieved this by building collective capacity for this work. Our findings offer theoretical and practical guidance for sustainably organizing work in occupations marked by high levels of emotionally charged work.

Links to Previous Issues

Issue Date	Vol(Iss.)	Link
July, 2026	48(3)	https://www.apadivisions.org/division-5/publications/score/2026/07/issue.pdf
April, 2026	48(2)	https://www.apadivisions.org/division-5/publications/score/2026/04/issue.pdf
January, 2026	48(1)	https://www.apadivisions.org/division-5/publications/score/2026/01/issue.pdf
October, 2025	47(4)	https://www.apadivisions.org/division-5/publications/score/2025/10/issue.pdf
July, 2025	47(3)	https://www.apadivisions.org/division-5/publications/score/2025/07/issue.pdf
April, 2025	47(2)	https://www.apadivisions.org/division-5/publications/score/2025/04/issue.pdf
January, 2025	47(1)	https://www.apadivisions.org/division-5/publications/score/2025/01/issue.pdf
October, 2024	46(4)	https://www.apadivisions.org/division-5/publications/score/2024/10/issue.pdf
July, 2024	46(3)	https://www.apadivisions.org/division-5/publications/score/2024/07/issue.pdf
April, 2024	46(2)	https://www.apadivisions.org/division-5/publications/score/2024/04/issue.pdf
January, 2024	46(1)	https://www.apadivisions.org/division-5/publications/score/2024/01/issue.pdf
October, 2023	45(4)	https://www.apadivisions.org/division-5/publications/score/2023/10/issue.pdf
July, 2023	45(3)	https://www.apadivisions.org/division-5/publications/score/2023/07/issue.pdf
April, 2023	45(2)	https://www.apadivisions.org/division-5/publications/score/2023/04/issue.pdf
January, 2023	45(1)	https://www.apadivisions.org/division-5/publications/score/2023/01/issue.pdf
October, 2022	44(4)	https://www.apadivisions.org/division-5/publications/score/2022/10
July, 2022	44(3)	https://www.apadivisions.org/division-5/publications/score/2022/07
April, 2022	44(2)	https://www.apadivisions.org/division-5/publications/score/2022/04
January, 2022	44(1)	https://www.apadivisions.org/division-5/publications/score/2022/01
October, 2021	43(4)	https://www.apadivisions.org/division-5/publications/score/2021/10
July, 2021	43(3)	https://www.apadivisions.org/division-5/publications/score/2021/07
April, 2021	43(2)	https://www.apadivisions.org/division-5/publications/score/2021/04
January, 2021	43(1)	https://www.apadivisions.org/division-5/publications/score/2021/01
October, 2020	42(4)	https://www.apadivisions.org/division-5/publications/score/2020/10
July, 2020	42(3)	https://www.apadivisions.org/division-5/publications/score/2020/07
April, 2020	42(2)	https://www.apadivisions.org/division-5/publications/score/2020/04
January, 2020	42(1)	https://www.apadivisions.org/division-5/publications/score/2020/01

Issues of The Score can be accessed at the following webpage:

<https://www.apadivisions.org/division-5/publications/score/issues>

Division 5 Membership Information



Division 5 draws together specialists in the areas of quantitative and qualitative methods. Membership includes specialists in the fields of educational psychology, industrial-organizational psychology, individual assessment, measurement, program evaluation, psychometrics, qualitative inquiry, research interviewing, research methods and statistics. Division 5 provides the opportunity to participate in this critical aspect of professional development, and is a vibrant community that serves the core interests and needs of rigorous scientific inquiry in the social sciences. Those who are interested in joining the division or who would like more information should be referred to the Division 5 website:

<https://www.apadivisions.org/division-5/membership>

Advertising in The Score

The Score is the newsletter of the American Psychological Association's Division 5—Evaluation, Measurement, and Statistics. Division 5 is concerned with promoting high standards in both research and practical application of psychological assessment, evaluation, measurement, and statistics. Approximately half of the Division 5 members are university faculty members in quantitative psychology, psychometrics, educational psychology, or industrial organizational psychology and half are engaged in careers in industry, including the areas of individual and large-scale assessment. More than 1,000 Division 5 members receive The Score each quarter. Advertisements in The Score may be in the form of display advertisements or job announcements. Both types of ads can include graphics and other design features and can be submitted as text or camera-ready display art. Prices for advertisements and size requirements are provided in the accompanying table. Submission deadlines are 45 days prior to publication: February 15 for the April issue, May 15 for July, August 15 for October, and November 15 for January.

To advertise in The Score, please contact Editor Kevin Williams at kmwilliams@ets.org.

Size	Dimensions	Display Ad Price	Job Announcement Price
Full page	7.125" × 9.5"	\$235	Not available
Half page	7.125" × 4.75"	\$175	\$120
Third page	4.75" × 4.75"	\$125	\$90
	2.375" × 9.5"		
Sixth page	2.375" × 4.75"	\$90	Free/\$55
Notes: Insertion orders for four consecutive issues receive a 15% discount. First sixth page job ad each year free, thereafter \$55.			

Publishing in The Score

The Score publishes ~1,000-word summaries on topics relevant to quantitative statistical methods, research methods, qualitative methods, assessment, and/or education. Submissions may be original research or a summary of research disseminated elsewhere (conference presentations, journal articles, etc.). There are no publication fees. For more information or to contribute your research to The Score, please contact Editor Kevin Williams (kmwilliams@ets.org). Submissions are accepted in a priority sequence based on when they are received.