WHITE PAPER

Leveraging sentiment analysis to evaluate clinical outcome assessments

A KEY QUESTION

If sentiment improves as engagement drops, what does that mean for your endpoint strategy?

KEYWORDS Sentiment Analysis, Clinical Outcome Assessments, Cognitive Debriefing, Patient-Reported Outcome, Content Validity

Contributors:
Christine Bradshaw, MA, Senior Research Associate, Patient-Centered Endpoints, Fortrea
Kristina Davis, PhD, Director, Patient-Centered Endpoints, Fortrea
Kelly Stout, Research Associate, Patient-Centered Endpoints, Fortrea

Clinical Outcome Assessments (COAs) are measures designed to provide insight into a patient's disease or condition by describing or reflecting on how a patient feels or functions. The scores produced by these measures support evaluations of effectiveness, dose optimization, safety and tolerability, to help determine the clinical benefits and risks of a medical product in clinical trials.[1] Given their central role in informing benefit-risk assessments in clinical trials, it is crucial that COAs accurately capture meaningful changes in the patient experience.

The most common method for evaluating COA content is cognitive debriefing, a qualitative interview approach conducted with patients to confirm the relevance, clarity and interpretability of items and response options. Cognitive debriefing interviews typically include standardized probes such as, *What does this question mean to you?, Was this question easy or difficult to answer?, How did you decide on your response?,* to elicit detailed feedback on items in support of item-level evaluation of content validity.[1]

Given the importance of understanding patients' perceptions of COAs, this study sought to explore whether additional analytic approaches could enhance interpretation of cognitive debriefing interview data. Specifically, we examined the use of sentiment analysis, an analytic technique that classifies the emotional tone of text into categories such as positive, negative or neutral.[2] Sentiment analysis may provide unique, complementary insight into patients' perceptions of COAs, particularly with respect to how patients feel about the items and response options.

Fortrea

Methods

To test whether sentiment analysis can support the assessment of feedback on COAs, we used 16 transcripts from a cognitive debriefing study on a Patient-Reported Outcome (PRO) measure (a type of COA that is a self-assessment of a patient's condition) for COVID-19 symptoms. The PRO is composed of 14 items; however, we limited our analysis to ten items that shared the same response options. This decision allowed us to minimize variability and maintain comparability across items.

Data processing

Transcripts were segmented into sentences and exported to Excel. Mentions of contextual details unrelated to feedback on the measure, such as descriptions of a participant's COVID-19 experience (e.g., I was quarantined for 15 days), were excluded from analysis to avoid distorting sentiment of the items. Two coders trained in qualitative analysis independently reviewed the data using a codebook that categorized sentiment as positive, negative or neutral:

- **Positive sentiment** indicated clear understanding of the item or instructions, ease of response or agreement with the response options, a perceived relevance of the item or liking other aspects of the item *(e.g., This was easy for me to answer)*
- **Negative sentiment** indicated a lack of understanding of the item or instructions, difficulty responding or disagreement with the response options, a perceived lack of relevance of the item or disliking other aspects of the item *(e.g., I had to reread this a few times to understand)*
- **Neutral sentiment** was applied to feedback that was ambiguous, lacked emotional tone or was factual without evaluative language *(e.g., This question is about general fatigue)*

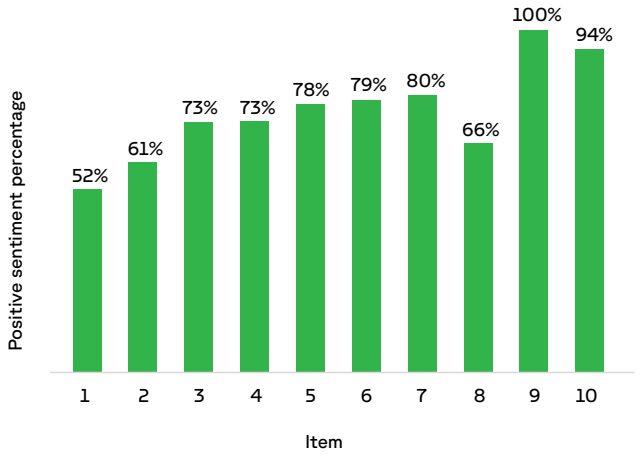
First, frequency counts for the number of positive, negative and neutral statements were calculated in R and converted into percentages for each item to calculate item sentiment. Initial examination of these descriptive results revealed a pattern in which positive sentiment appeared to increase across the sequence of items debriefed during the interview. Based on these initial results, we hypothesized that this pattern may be attributed to satisficing, the tendency for respondents to use less cognitive effort over time by providing shorter, less reflective or more uniformly positive responses.[3] To evaluate this hypothesis, we conducted Spearman's rank order correlation analyses in R to assess the strength and direction of the relationship between the following ranked variables:

- **Item order:** Defined as the sequence in which items were debriefed during the interview (ranked from 1 = first item to 10 = last item)
- **Word count:** Defined as the number of words spoken by the participant that were directly related to interpreting, understanding or responding to an item, excluding unrelated conversational remarks (ranked from 1 = fewest words to 10 = most words)
- **Positive sentiment percentage:** Defined as the proportion of statements coded as positive for each item (ranked from 1 = lowest to 10 = highest)

Results

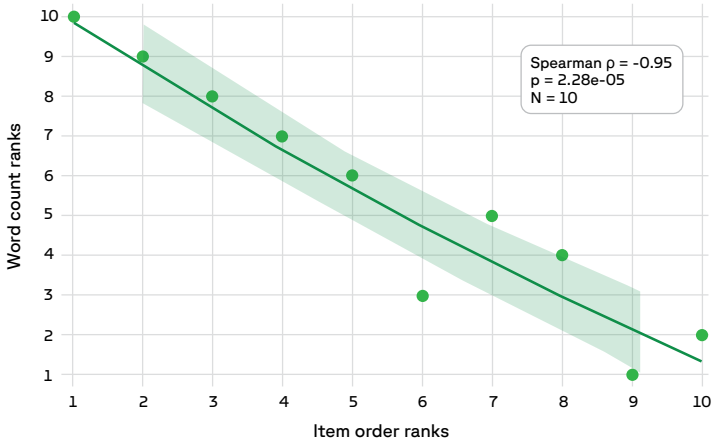
Positive sentiment frequencies by item

Descriptive analysis of frequencies indicated that all ten items exhibited positive sentiment percentages exceeding 50%. Additionally, positive sentiment appeared to increase across the sequence of items debriefed during the interview.



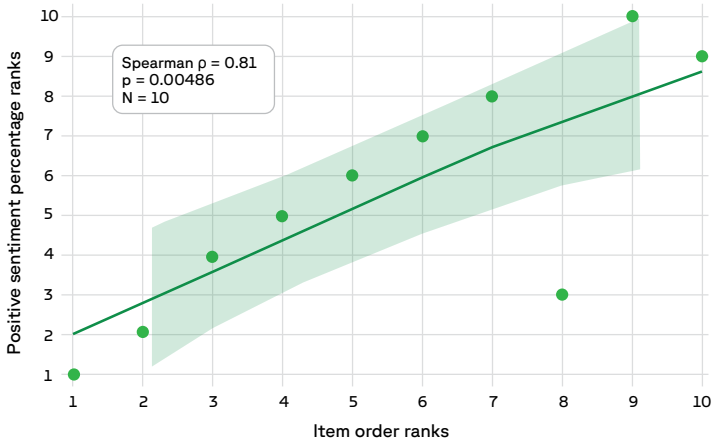
Item order vs. word count

A very strong negative one-directional relationship (ρ = -0.9515, p < 2.2e-16) was observed between item order and word count, indicating that participants provided more feedback on items at the beginning of the interview compared to the end.



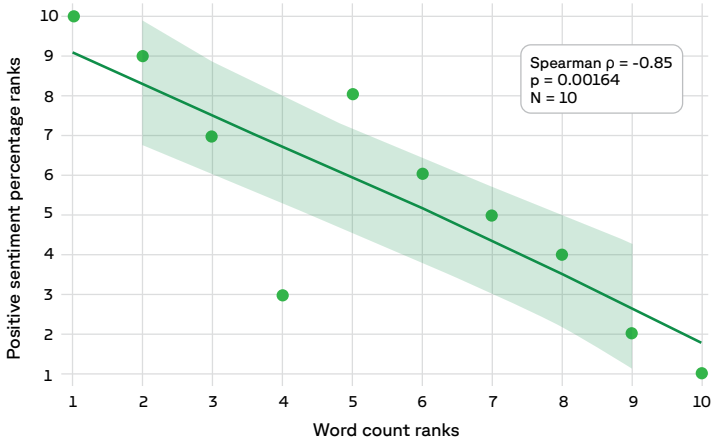
Item order vs. positive sentiment percentage

A very strong one-directional relationship (ρ = +0.84499, p = 0.00824) was observed between Item order and positive sentiment percentage, indicating that later items were associated with higher positive sentiment percentage.



Word count vs. positive sentiment percentage

A very strong negative one-directional relationship (ρ = -0.85, p = 0.00164) was observed between word count and positive sentiment percentage, indicating that as respondents' answers grew shorter, their sentiment became more positive.



Findings

Sentiment analysis revealed patterns that may inform both content validity evaluation and the cognitive debriefing process of COAs. Across all items, positive sentiment exceeded 50%, indicating generally favorable feelings towards the measure and its items. Participants' positive feelings regarding item and instruction clarity, relevance and ease of response may provide complementary evidence to support the measure's content validity by capturing the emotional valence of interpretations beyond traditional cognitive debriefing. However, positive sentiment percentages must be interpreted cautiously in the presence of satisficing behavior.

Spearman's results revealed a steady increase in positive sentiment percentage and a corresponding decrease in word count across item order. These results indicate that earlier items elicited longer, more detailed and more critical feedback, whereas later items produced shorter and more agreeable responses with higher positive sentiment. The observed inverse relationship between positive sentiment percentage and word count further supports this pattern.

Future directions

The sentiment expressed by participants about a COA and its items may strengthen evidence for content validity at both item and measure levels.

Sentiment score frequencies may be incorporated into item-level decisions, such as identifying items for revision or removal. For example, items lower in positive sentiment—or higher in negative or neutral sentiment—may indicate issues with clarity, relevance or interpretability and should be prioritized for qualitative follow-up.

Sentiment analysis also shows promise as a tool for detecting satisficing during cognitive debriefing interviews through the examination of positive sentiment frequencies and Spearman's correlation patterns. Given the evidence of satisficing during this cognitive debriefing, randomizing item order may be key to mitigating possible order effects.[3] We also recommend implementing more engagement monitoring throughout the interview, such as engagement checks (e.g., restatements, probes) to help maintain processing depth for all questions. Evaluating these design modifications will help determine whether item order effects are driven primarily by the items themselves or by declining participant engagement over time.

Future efforts can work towards automating sentiment analysis using tools such as the sentimentr package in R to reduce coder burden. Establishing agreement between automated and human-coded sentiment would clarify whether an automated approach is reliable enough to supplement qualitative analysis.

Limitations

Our findings are based on a single PRO measure and a relatively small sample of transcripts (n = 16), which may limit generalizability. Segmenting transcripts for coding may have compromised coding accuracy by removing essential context. Additionally, the observed Spearman's correlations do not establish causation between item order, word count and positive sentiment percentage. Although Spearman's correlation can be computed with small samples, our sample size (n = 10) provides limited power and wide uncertainty around the estimates. Therefore, these findings should be interpreted as exploratory.

Conclusions

Across ten items and 16 cognitive debriefing interviews, sentiment analysis revealed strong one-directional relationships among item order, word count and positive sentiment percentage. Earlier items elicited longer and more critical feedback, whereas later items produced shorter and more agreeable responses, consistent with satisficing behavior.

Although positive sentiment frequencies reflected generally favorable perceptions of item and instruction clarity, relevance, ease of response and agreement with response options, these findings show that positive sentiment must be interpreted in the context of participant engagement and item order. While not a replacement for qualitative interpretation, sentiment analysis may serve as a useful supplementary tool for identifying order effects and engagement-related patterns that could influence content validity conclusions and item-level decision making for COAs. However, we encourage researchers to recognize the limitations of this work and take steps to address them in future studies.

Learn more about our approach to Patient-Centered Endpoints.

Connect at fortrea.com

References

1. Patient-focused drug development: Selecting, developing or modifying fit-for-purpose clinical outcomes assessments. *FDA*, 2025. https://www.fda.gov/media/159500/download
2. Pang B, Lee L. Opinion mining and sentiment analysis. *Now Publishers*, 2008. doi:10.1561/9781601981516
3. Krosnick JA, Alwin DF. An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Q.* 1987;51(2):201-219. doi:10.1086/269029

Fortrea