International Journal of Economics, Business and Management Research

Vol. 6, No.02; 2022

ISSN: 2456-7760

Validity of AI-generated One-off Questionnaires for Organizational Transformation

Dr. Jan van de Poll[1], Yang Yong[2], Nicole van de Weijer[3]

[1] Managing Director at Transparency Lab BV (www.praioritize.com)

[2] Chief Technology Officer at Transparency Lab BV

[3] Algorithm Engineer at Transparency Lab BV

doi: 10.51505/IJEBMR.2022.6208 URL: http://dx.doi.org/10.51505/IJEBMR.2022.6208

Abstract

Strategic decision-making is a precise craft that usually happens under time pressure. And often, the data for such decision-making is not (enough) available in the corporate data warehouse: you have to ask employees. This employee polling is usually very time-consuming**.** In this study, a novel approach is developed that couples artificial intelligence (AI) and a specific survey scale format to make one-off questionnaires in near real-time. We tested such AI-generated one-off questionnaires in 23 strategic situations where almost 7,000 employees provided nearly 6 million answers. Six statistical parameters assess the validity and reliability of the questionnaires. Our test results reveal that the developed methodology saves time and produces valid survey outcomes. Our finding, as a rule-of-thumb, is that, above a sample size of 100 respondents, AI-generated one-off questionnaires are scoring well on the selected validity/reliability parameters. Consequently, the developed technique could be employed successfully in generating valid and reliable one-off questionnaires for organizational transformation.

Keywords: Employee polling, organizational transformation, Guttman-Poll, GPT-3

Introduction

In almost every organization, there is a strategy on how the company should proceed in the near future. In larger organizations, such a strategy might even drive something like an ‘organizational transformation’ program affecting many, if not all, employees. In such a transformation, polling the views from many employees – the 'wisdom of the crowd' – has shown a valuable contribution to the decision-making process (Giles, 2005; Surowiecki, 2005).

In such polls, management often cares less about the employees’ feelings and opinions about the transformation. The organization’s management would like to tally how the transformation moves along objectively. We postulate that an organization’s strategy is by nature different from any other organization. Consequently, no generic questionnaire sufficiently assesses how far an organizational transformation for that particular organization has progressed. But if upper management starts to make a one-off, made-to-measure questionnaire themselves – or asks consultants to do so – how could they be sure they are building a construct that meets validity requirements? A strategy is about moving the organization to a new, future state. A state with which management might be partly unfamiliar. And if specialized consultants are not at the scene (e.g., they are deemed too expensive or the strategy is still too confidential to share), how

Vol. 6, No.02; 2022

ISSN: 2456-7760

does management ensure they have captured all the aspects of their transformation? They do have a policy document but might have forgotten certain strategic aspects. Therefore, we aim to employ an Artificial Intelligence (AI) technique that generates one-off questionnaires for organizational transformation that are fast and cheap to produce, relevant to the transformation issue at hand, and statistically valid and reliable.

A new breed of artificial intelligence in Natural Language Processing are generative pre-trained transformers that require very little training (in the form of text input) to create text output (so-called 'few shot learners'; Brown et al., 2020). OpenAI's GPT-3 application (https://www.openai.com/) can be trained to supply management with a potentially infinite number of topics and questions about a particular strategic issue. It's up to management to decide which topics best reflect their strategy or the aspects of their organizational transformation they most care about tracking. GPT-3 proposes, and management decides which issues and questions to keep or skip. Figure 1 shows a GPT-3 prompt (management's concern was about the organization's digital maturity) and a first few samples of AI-generated questionnaire topics (plus a brief explanation) about digital maturity. Not visible in Figure 1 are approximately 500 words of training data that only needed to be supplied once. This training data tells GPT-3 the text *structure* it should return, not the *content*. The prompt (visible in bold in Figure 1) determines the content to which GPT-3 has to relate.

KEYWORDS:
digital transformation, change, corporate reinvention, digitalization, employees, strategy, leadership, processes

SUGGESTED QUESTIONNAIRE TITLE:
Digital Transformation Monitor

QUESTIONNAIRE TOPICS:
1. Vision: where does digital fit in
2. Strategy: where is innovation needed
3. Leadership: inform and encourage employees
4. Communication: link initiatives to not reinvent the wheel
5. People: how to build a digital organization
6. Processes: how does digital change processes
7. Climate: how can people thrive and think outside the box
8. Technology: how digital can be used today
9. Funding: how to pay for it all
10. Culture: how do you upgrade your culture

Figure 1. AI-generated topic classifier: prompt/keywords (in bold), and resulting title, topics, and a brief explanation per topic.

How will such an AI-generated one-off questionnaire on organizational transformation perform in terms of validity? In the classical model of test validity, there are three types of validity: construct-, content-, and criterion validity (Guion, 1980; Brown 1996). Others consider construct validity the overarching concern of validity research (Messick, 1995) or assume that a questionnaire is valid if the attributes exist and attribute variations result in variations in the

Vol. 6, No.02; 2022

ISSN: 2456-7760

measurement outcomes (Borsboom, Mellenbergh, & Van Heerden, 2004). Related to validity is reliability and refers to the consistency of a measure: over time (test-retest reliability), across items (internal consistency), and across different researchers (inter-rater reliability) (Shaughnessy, Zechmeister, & Zechmeister, 2000). The main objective of the current study is to test a new methodology developed in this study to generate valid, one-off questionnaires accurately. Therefore, we aim to answer the following research questions:

1) How do A.I.-generated one-off questionnaires about aspects of organizational transformation perform in terms of validity?
2) Are there general statements to make about their validity?
3) Must certain precautions be taken to make such statements?

Method

Procedure and participants

When an organizational transformation is underway, management would like to measure the transformation's progress objectively. The ship has sailed concerning employees' feelings and opinions about the strategy. Hence, we first forewent a survey using Likert scales to tally measure verifiable facts or behavior objectively. Therefore, we designed an alternative survey scale based on the Guttman scale (Stauffer et al., 1950; Diamond, McDonald, and Shah, 1986) better geared to objectively polling employees (Van de Poll 2018, 2021, and very recently, Van de Poll et al., 2022).

Next, we analyzed 23 different employee polls about various strategic issues, all requiring some organizational transformation, which included topics on - among others - employee engagement, innovation, work processes, competencies, digital transformation, work pressure, technology adoption, team effectiveness, and IT security. These employee polls showed a response from 6,912 respondents in 726 teams, giving 5,985,792 answers. The number of employees per poll ranged from 13 to 957. We used PRAIORITIZE, an automated consultancy platform **(www.praioritize.com)**, which generates one-off questionnaires using AI to apply organizational transformations.

Measures

Our alternative survey format based on the Guttman scale is an ordinal, multiple-choice scale where every following answer is better than the answer before. Uhlaner (2002) calls these 'breaking points.' For example (from a team effectiveness poll):

Q. How do you celebrate successes?

1. We don't
2. When there is an apparent reason to do so, with whoever is involved
3. We make it a habit to celebrate successes with the entire team

We considered this question format verifiable (Ahrens & Chapman, 2006; Plewis & Mason, 2007). We abstained from adjectives or adverbs that couldn't be verified (e.g., "good"). We

Vol. 6, No.02; 2022

ISSN: 2456-7760

reduced the respondents' self-reporting bias (Donaldson and Grans-Vallone, 2002) by adding "proof-words" like, e.g., 'formally,' 'measurable,' 'periodically,' 'described' and 'documented.' Such "proof-words" reduce the emotional or cognitive meaning given by employees to the answers (Frese & Zapf, 1988). We summarized the most used aspects of validity and reliability in Table 1. We based this table partially on Onwuegbuzie et al. (2007).

Table 1
Indicators of validity and reliability (cf. Onwuegbuzie et al., 2007)

Indicator	Explains	Applicability	Validation
Content-related validity			
Face validity	Questions are relevant to respondent	Management designs questionnaire	--
Item validity	Questions represent the intended area	Policy document as the source	% Extreme response styles
Sampling validity	Questions cover all aspects of that area	Management prioritizes topics	Kaiser-Meyer-Olkin
Criterion-related validity			
Concurrent validity	Test outcomes correlate with earlier tests	Not applicable	--
Predictive validity	Test predicts an outcome	Longitudinal data not available	--
Construct-related validity and reliability			
Substantive validity	Questions represent the intended area	Policy document as the source	Cronbach's α, GL2, St.Cr.α
Structural validity	Scores reflect construct dimensionality	Questions evenly spread among topics	Cronbach's α, GL2, St.Cr.α
Outcome validity	Test produces consistent findings	Respondents score independently	GL4 split half correlation
Generalizability	Test useful for a broader respondent group	Expand to the entire organization	GL4 split half correlation
Comparative validity			
Convergent validity	Two tests relate in theory and practice	Not applicable	--
Discriminant validity	Two tests do not relate in theory and practice	Not applicable	--
Divergent validity	Two tests do not relate in theory and practice	Not applicable	--

GL2 = Guttman Lamba 2, St.Cr.α = Standarddized Cronbach , GL4 = Guttman Lambda 4

The applicability of validity and reliability indicators to A.I.-generated one-off questionnaires for organizational transformation have been summarized in the column 'Applicability.' For face validity, we deem management capable enough to determine which topics and questions represent the transformation at hand.

We could say the same for item validity, but choosing topics and phrasing of questions might encourage employees to resort to so-called 'extreme response styles.' An 'acquiescence response style' (De Beuckelaer, Weijters, & Rutten, 2009) happens when respondents choose (almost) everywhere the best answer. Another extreme response style happens when respondents frequently choose the middle answer in the questionnaire scale. Hence, we totaled the percentage of employees where 90% of the responses out of the three multiple-choice options were identical in the column Extreme Response Styles (% ERS) in Table 1.

For sampling validity, we reckon that management can determine the priority topics for polling but deem a Kaiser-Meyer-Olkin (KMO) score as an indicator for sampling adequacy.

Looking at criterion validity, we postulate that concurrent validity is not applicable for a one-off questionnaire.

Vol. 6, No.02; 2022

ISSN: 2456-7760

The predictive validity would undoubtedly be interesting to measure when multiple measurements of the same one-off questionnaire were available.

For construct validity and reliability, it's a matter of how much questions and respondents correlate with each other.

Generalizability is essential when only a part of the employees respond: can the outcomes relate to the entire organization?

We apply four measures for this construct validity and reliability section: Cronbach's alpha (CrA), Guttman's L2 (GL2), Standardized Cronbach's alpha (St.Cr.A, to cater for questions that might have a different number of multiple-choice answers) and the split-half correlation in a Guttman's L4 Split Half mode (GL4SH) analysis. We refrained from further analysis concerning comparative validity due to the one-off nature of the questionnaires.

Data analysis

Table 2 describes the 23 assessments, including the topic, the number of questions, and the number of participating teams and employees. The table also shows the scores of our six validation and reliability checks. Table 2 is sorted by the number of respondents (indicated by 'Sort ↓'). When the six parameters are scoring below certain quality thresholds (indicated in Table 2's footer), the parameter is highlighted in bold.

The right-most column in Table 2 shows how many of the six parameters have been highlighted in bold: a validity/reliability verdict, if you will.

The item validity is represented by the percentage of respondents with % ERS. We postulate a percentage of two to three percent of our cut-off between good and bad, although various authors have reported double-digit %ERS (e.g., Liu et al., 2017).

The sampling validity is covered by de KMO score. A KMO score of 0.70 to 0.80 is considered 'good,' from 0.80 to 0.90 'great,' and above 0.90 'superb' (Streiner and Norman, 2015).

The construct validity/reliability section is analyzed using CrA, GL2, St.Cr.A and the split-half correlation in a GL4SH analysis. A CrA score should be at least 0.70, provided that the number of questions is clearly above 10 (Nunnally, 1978 GL2 should preferably be higher than 0.70 for group evaluations (Callender & Osburn, 1979). Guttman's L4 should preferably be above 0.85 (Benton, 2015), but as we added the Split Half option (GL4SH), we prefer the L4 to be as low as possible and set a boundary of (plus or minus) 0,15.

Table 2
Sample size and selected validity and reliability indicators

Transformation topic	#.Q	#.T	#.E	Item %ERS	Sample KMO	Construct / Reliability CrA	GL2	StCrA	GL4SH	Items in bold
Sort ↓										
Assessments sorted by the number of employees										
Marketing strategy	25	1	13	0.0%	**0.471**	**0.526**	**0.658**	**0.508**	**-0.414**	5
Managing employee well-being	31	2	22	0.0%	**0.337**	0.806	0.837	**0.792**	**-0.507**	3
ICT deployment	79	1	23	0.0%	**0.524**	0.805	0.848	**0.743**	*N.A.*	2
Business requirements ERP	44	8	31	0.0%	**0.364**	0.829	0.854	0.843	**-0.491**	2
Strategy implementation (municipality)	40	20	49	0.0%	**0.453**	0.935	0.941	0.935	**-0.232**	2
Team effectiveness	28	3	50	0.0%	**0.559**	0.858	0.873	0.866	-0.126	1
Digital Maturity Scan	10	4	74	0.1%	0.743	0.715	0.747	**0.692**	0.134	1
Factory technology deployment	38	43	75	0.0%	**0.485**	0.776	0.806	**0.789**	-0.107	2
Corporate Commuinications Monitor	30	8	86	0.0%	0.799	0.902	0.911	0.904	0.119	
Strategy deployment in SME firms	38	*N.A.*	106	0.1%	0.741	0.900	0.905	0.902	0.001	
ICT service assessment	34	7	119	0.0%	**0.644**	0.830	0.840	0.832	-0.008	1
Controller competencies	37	*N.A.*	228	1.3%	0.906	0.939	0.940	0.939	-0.081	
Staff department evaluation	40	20	268	0.2%	0.818	0.884	0.905	0.886	-0.045	
Strategy implementation (care)	37	21	334	0.0%	0.882	0.921	0.924	0.923	0.085	
Leadership assessment	40	74	433	0.0%	0.878	0.875	0.880	0.881	-0.121	
ERP implementation	33	52	504	0.0%	0.746	0.786	0.797	**0.791**	0.039	1
Strategy implementation (insurer)	58	18	518	2.4%	0.974	0.976	0.977	0.976	0.040	
Employee engagement survey	29	12	550	0.1%	0.881	0.872	0.878	0.872	-0.050	
Working under Covid-19	29	78	587	0.0%	0.724	0.723	0.740	**0.741**	0.008	1
IT services performance	39	126	598	0.0%	0.882	0.882	0.888	0.878	-0.086	
Employee technology adoption	42	7	640	0.1%	0.874	0.878	0.882	0.878	0.074	
Employee work pressure	43	91	647	0.0%	0.780	0.813	0.826	0.804	0.012	
Process adoption in the IT department	42	130	957	0.7%	0.914	0.911	0.914	0.912	0.010	
Total										
Total	866	726	6,912							
Average	38	35	301							
Weighted average				0.19%						

#.Q = Number of questions, #.T = Number of teams, #.E = Number of employees
*KMO = Kaiser-Meyer-Olkin (**below 0.7** highlighted **in bold**), %ERS = Percentage respondents with Extreme Response Styles*
*CrA = Cronbach's Alpha (**below 0.7** highlighted **in bold**), GL2 = Guttman L2 (**below 0.7** highlighted **in bold**)*
*StCrA = Standardized Cronbach's alpha (**below 0.8** highlighted **in bold**)*
*GL4SH = the Split Half correlation in a Guttman L4 in Split Half mode analysis (**above** +/- **0.15** highlighted **in bold**)*

Vol. 6, No.02; 2022

ISSN: 2456-7760

Results

Table 2 also shows how the questionnaires scored on the six criteria for item-, sample, and construct validity/reliability. We assume the questionnaire's 'one-offness' to cater to the face validity. Simultaneously, this 'one-offness' renders concurrent, predictive, convergent, discriminant, and divergent validity not relevant for our situation. For example, concurrent validity indicates whether earlier outcomes concur with the current test. In a one-off situation, there are no earlier outcomes with which to compare.

The percentage of extreme response styles seems independent of the number of respondents. Otherwise, the higher the number of respondents, the more the remaining validation criteria indicate validity and reliability. Below 100 respondents, the criteria still flip-flop between good and not so good. Roughly above 100 respondents, the criteria confirm the questionnaire's validity and reliability. We deem the 0,19% of respondents with extreme response styles more an attribute of the Guttman-Poll scale than the result of using artificial intelligence. The dotted line, halfway Table 2, indicates where the number of 'bold' parameters are dropping to (almost) zero when raising the number of respondents. The assessment below the dotted line has 86 respondents. Beware of the risk of overfitting our model, we conclude that, as a rule-of-thumb, minimally a hundred respondents are likely to be sufficient to ensure the questionnaires' validity and reliability. That said, a few assessments with more than a hundred assessments show one of the six parameters still in bold. That leads to the conclusion that the rule-of-thumb of one hundred respondents doesn't mean the person managing the assessment can refrain from any further statistical checks.

Limitations and future research

There are a few cautionary remarks to be made about our research. We have tried to vary the topics of the strategic assessments. Yet, that variety is just a first step in indicating validity and reliability. More variety in strategic issues, combined with a higher number of strategic assessments than the 23 in our sample, will confirm whether our "hundred respondents" cut-off value will stand. Additionally, more strategic assessments will also confirm whether the 0,19% of respondents with extreme response styles is a hard number to count on.

Conclusions

In today's frantic business environment, it's almost mandatory to very quickly assess the status of strategic projects and organizational transformations. Unfortunately, there is often no time for - and often no availability of - tried and true questionnaires. In addition, an organization's strategy is different from one organization to another. No generic questionnaire assesses in sufficient detail how far a particular organization has made progress. Therefore, producing tailor-made questionnaires in near real-time is critical to very quickly assess the status of strategic projects and organizational transformations. In this study, we presented a new approach that couples artificial intelligence and a specific survey scale format to generate statistically valid, one-off questionnaires in near real-time**.** The developed model is verified using six validation and reliability checks for 100^{+} respondents. We believe that this approach in this study may play a significant role in organizational transformations under time pressure.

Vol. 6, No.02; 2022

ISSN: 2456-7760

Acknowledgement

We would like to thank Dr. Jasna Duricic for her constructive comments during the planning and development of this research.

References

Ahrens, T., & Chapman, C. S. (2006). Doing qualitative field research in management accounting - positioning data to contribute to theory. Accounting, Organizations and Society, 31, 819-841.

Benton, T. (2015). An empirical assessment of Guttman's Lambda 4 reliability coefficient. In Quantitative psychology research (pp. 301-310). Springer, Cham.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. Psychological review, 111(4), 1061.

Brown, J. D. (1996). Testing in language programs. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Callender, J., & Osburn, H. (1979). An Empirical Comparison of Coefficient Alpha, Guttman's Lambda - 2, and MSPLIT Maximized Split-Half Reliability Estimates. Journal of Educational Measurement, 16(2), 89-99.

De Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. Quality & Quantity, 44(4), 761-775.

Diamond, I. D., McDonald, J.W., & Shah, I.H. (1986). Proportional hazards models for current status data: application to the study of age at weaning differentials in Pakistan. Demography 23(4), 607-620. DOI: 10.2307/2061354

Donaldson, S. I., & Grans-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. Journal of Business and Psychology 17(2), 245-260.

Frese, M., & Zapf, D. (1988). Methodological issues in the study of work stress: Objective vs subjective measurement of work stress and the question of longitudinal studies. In: C. L. Cooper, & R. Payne (Eds.), Causes, Coping, and Consequences of Stress at Work (pp. 375-411). Wiley & Sons, Chichester.

Giles, J. (2005). Wisdom of the crowd. Nature 418(17), 281-281.

Guion, R. M. (1980). "On trinitarian doctrines of validity". Professional Psychology. 11 (3): 385–398.

Guttman, L. (1950). Chs. 2, 3, 6, 8 and 9 in SA Stauffer, L. Guttman, EA Suchman, PF Lazarsfeld, SA Star and JA Clausen. 1950. Studies in Social Psychology in World War II, 4, 46-90.

Liu M, Harbaugh AG, Harring JR and Hancock GR (2017) The Effect of Extreme Response and Non-extreme Response Styles on Testing Measurement Invariance. Front. Psychol. 8:726

Vol. 6, No.02; 2022

ISSN: 2456-7760

Messick, S. (1995). "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning". American Psychologist. 50 (9).

Nunnally, J. C. (1978). An overview of psychological measurement. Clinical diagnosis of mental disorders, 97-146.

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. American Educational Research Journal, 44(1), 113-160.

Plewis, I., & Mason, P. (2007). What works and why: combining quantitative and qualitative approaches in large-scale evaluations. International Journal of Social Research Methodology, 8(3), 185-194.

Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2000). Research methods in psychology. McGraw-Hill.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). Health measurement scales: a practical guide to their development and use. Oxford University Press, USA.

Surowiecki, J. (2005). The Wisdom of the Crowds. Anchor, New York.

Uhlaner, L. M. (2002). The use of the Guttman scale in development of a family business index. (No. H200203). EIM Business and Policy Research.

Trochim, W. (2002). Threats to construct validity and Pattern Matching for Construct Validity. Available at@ p.

Van de Poll, JM (2018) Ambition patterns in strategic decision-making, Doctor of Philosophy, Industrial Engineering and Innovation Sciences, Technical University Eindhoven.

Van de Poll, JM (2021). An Alternative to the Likert Scale When Polling Employees. International Journal of Business and Management, 9(5), 239-244.

Van de Poll, JM, Shamsi, A., Brouwer, A., Miller, M., (2022) Operationalizing purpose between the actual situation and ambition, International Journal of Business Management and Economic Review, Vol (5), 9-20.