DOI: 10.19181/4m.2023.32.2.4
EDN: PWCWQK

GATA: TEST – RETEST RELIABILITY OF MEASUREMENT OUTCOMES

Oleg L. Chernozub
Institute of Sociology FCTAS RAS, Moscow, Russia
9166908616@mail.ru
ORCID: 0000-0001-5689-8719

For citation: Chernozub O. L. GATA: test – retest reliability of measurement outcomes. *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2023, no. 57, p. 112–137. DOI: 10.19181/4m.2023.32.2.4.

Social researchers have long sought to overcome the vulnerability of the self-report method to a variety of effects based on respondents' inability or unwillingness to answer correctly. One obvious strategy for solving this problem is to try to extract information in such a way that the respondent's attitude towards the object under study can be assessed by his or her actions "indirectly" related to that object, without relying on the respondent's "direct" answers as to what he or she thinks that attitude is. One of the many tools that claim to be able to probe the respondent's consciousness in this way is the Graphical Associative Test of Attitudes (GATA).

This paper presents a summary of the test-retest reliability of GATA. 18 repeat tests two (12 tests) and four (6 tests) weeks after the initial measurement showed that GATA can potentially produce reliable results. At the same time, this ability is not completely stable, it depends on the subject being assessed and the time between the tests being compared. While the reliability of GATA over a two-week period is often satisfactory (Spearman $\rho > 0.700$ in 4 out of 12 tests) and comparable to that of the self-report method, over a four-week period, only one out of six GATA evaluations maintains such a high level of reliability.

Possible dimensions of future research aimed at improving the reliability of GATA output are discussed in the paper.

Keywords: self-report; direct measurements; indirect measurements; test-retest reliability; GATA.

1. The aim and scope of this contribution

Attitudes are thought to influence people's behavior in alignment with those attitudes: a positive or negative attitude toward an object [1] leads to an approach or avoidance of that object [2]. Therefore, to predict people's behavior, it is essential to measure their attitudes [3]. However, "direct" measurement of attitudes encounters several challenges [4]. The self-reporting method, which currently dominates social research, has two significant drawbacks: respondents may be unable or unwilling to fully express their true attitudes, some of which may remain unrecognized by both the researcher and the respondent [5]. One way to address these issues is by supplementing "direct" self-report measures with "indirect" ones [6].

A measurement is considered "indirect" if it avoids the process of self-assessment or self-translation of attitudes [7]. Typically, the attitudinal object is presented, but the researcher does not ask participants to report their attitudes or preferences toward it. In some cases, the researcher may even ask respondents to avoid being influenced. Nonetheless, it is reasonably expected that spontaneous preferences will still influence certain behavioral aspects being measured. Currently, there are a number of instruments available to measure attitudes "indirectly" (see Appendix A).

The key criterion for "indirect" measurement is its capacity to reflect "mental content" regardless of the respondent's intentions or efforts to express or even conceal it. An opportunity to overcome the limitations of "direct" measurement arises when the measurement results are generated unintentionally and remain beyond the respondent's control [8].

This paper aims to evaluate the test-retest reliability of the Graphical Association Test of Attitudes (GATA), an "indirect"

attitude measure. Since its introduction in 2015, GATA has been incorporated into numerous predictive models of electoral, consumer, and communication behavior, demonstrating its effectiveness as an incremental factor in predictive accuracy (for an overview of forecasting practices, see [9]; for a meta-analysis of 64 cases, see [10]). However, despite its broad and successful practical use, GATA lacks formal validation.

Thus, this article seeks to address certain methodological and instrumental issues concerning GATA's application as a tool for "indirect" attitude measurement. Specifically, we aim to test its reliability by assessing the reproducibility of its output data over time.

We do not test any particular theoretical assumptions of GATA in this study, nor do we expect direct theoretical implications. Instead, the following sections present the main results of a large-scale experiment designed to test the reliability of GATA measurements, evaluated by their consistency over time. The focus of these experimental measurements is attitudes toward objects, understood as positive or negative evaluations of these objects.

2. GATA

The Graphical Association Test of Attitudes (GATA) intentionally avoids respondents' direct assessment of their attitudes toward the objects under investigation and, therefore, qualifies as a fundamentally "indirect" instrument. GATA was introduced as a supplementary measurement technique to complement the common self-report method [11]. Given the well-known limitations of self-reporting, we hypothesize that the accuracy of behavioral prediction models based on it could be improved by incorporating the "indirect" measurement of attitudes. Incremental effects should arise from a comprehensive combination of "directly" and "indirectly" measured attitudes, which can add to and correct one another [12].

To achieve this goal, GATA uses a chain of two sequential associative procedures.

In the first step, a respondent is presented with a primary stimulus representing an object of interest, followed by a set of target stimuli represented as abstract graphical shapes (Figure 1). To mask the researcher's true objective, the primary stimulus is embedded within a series of distractor stimuli. The output of this first step is the graphical shape(s) that the respondent associates with the object under study.

A "diverting pause" follows, with exposure to unrelated stimuli—typically common self-report questions from non-GATA sections of thequestionnaire.

In the second step, a phrase containing verbal markers of the approach–avoidance tendency is presented as the primary stimulus. This phrase usually includes wording such as "would like to look at," "would be nice to have around," or "would like to touch," among others. The stimulus phrase is then followed by the same set of graphic shapes.

In both stages, the respondent's task is to select from the target stimuli the graphical shapes perceived as "similar" or "close to" the primary stimulus. In this way, GATA is designed to produce an "indirect" measurement outcome.

The procedure for this method is structured as follows:

a. The respondent familiarizes themselves with the object of study, presented as a verbal concept on the screen of a CAPI device. A set of graphic shapes is displayed on the screen, and the respondent associates these shapes with the test object.
b. The respondent's attention is then diverted to other survey questions, preferably unrelated to the subject under study.
c. The respondent responds to an approach–avoidance phrase, ranking the graphic shapes from most to least preferred for prolonged contact.
d. An "individual scale" of preferences for graphic shapes is created based on the ranking from phase "c."

e. The implicit preference score, according to this "individual scale," is assigned to the studied object based on the association from phase "a."

As a result, each tested object receives a score on an ordinal scale, independent of the specific shapes that individual respondents may prefer or dislike due to psychological, cultural, mental, physical, or other similar factors.

Figure 1. An example of the GATA set of graphical shapes

Thus, methodologically, GATA claims to be an "indirect" measurement technique capable of producing results that are additive to, or even orthogonal to, "direct" measurements.

3. The experiment

3.1. General design

To test the reliability of the GATA measurement using a test-retest procedure [13], we designed a questionnaire combining two types of indicators:

SR: Self-report "direct" questions on an 8-point Likert ordinal scale.

GATA: GATA procedure, also on an 8-point ordinal scale.

The SR indicator is intended to serve as a "direct" measurement instrument, while GATA acts as an "indirect" measurement instrument. Both aim to hypothetically indicate the same attitudes, with the SR indicator considered the "control" and the GATA indicator as the "experimental" measure. This additional indicator was established as a baseline for assessing the extent of "normal" deviation in measurement results over time for our effective samples and across

the evaluated objects, providing a framework to structure the field of investigation.

As study objects, we selected six potentially ambivalent behavioral patterns, drawn from the questionnaire of the 7th wave of the World Value Survey (WVS) project.

1. Suicide.
2. Execution.
3. Tax avoidance.
4. Corruption.
5. Divorce.
6. Domestic violence.

The wording of the questions was used as it appears in the Russian version of the WVS questionnaire[1]. For details, please see Appendix B.

The issue we aimed to address with our set of indicators was the test-retest reliability of scales (measures) assessing attitudes toward these behaviors. The study's test-retest procedure included a baseline test and two subsequent retests for both "direct" and "indirect" measurements. Both retest intervals were set at two weeks, based on the assumption that this is sufficient time to prevent respondents from mechanically recalling their previous selection of GATA stimuli.

The study's primary hypothesis was that GATA measurements are reproducible/reliable and not merely indicative of random measurement error. This was broken down into three technical hypotheses:

(**$H_0$1**): "The test-retest reliability of GATA is less than Spearman $\rho < 0.700$ for every object evaluated."

(**$H_0$2**): "The test-retest reliability of GATA maintains its initial two-week interval grade over a four-week interval."

(**$H_0$3**): "The test-retest reliability of GATA remains consistent within the same time interval across all evaluated objects."

[1] See: World Values Survey Wave 7 (2017–2022). URL: https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp (date of access: 30.09.2024).

3.2. The procedure

The survey was technically organized as an online interview. The sampling method used was river sampling, where respondents were invited to participate via popular internet sites. Although our samples did not claim to "represent" any population, quotas for gender, age, type of residence, and federal district were used to roughly control the final sample through the algorithms of the software employed. The stimuli were optimized for both desktop and smartphone screens.

In line with the questionnaire structure, respondents began with self-report questions, followed by a block of unrelated questions. The GATA procedure was positioned in the middle of the questionnaire, with the association and ranking tasks separated by another block of unrelated questions. All survey instruments were designed as a single stream of tasks, optimized for either computer or smartphone screens, and programmed for online administration.

At the end of the initial test and the first retest, we asked respondents for permission to contact them again for the next stage of the survey. We only re-contacted those who provided such permission. These respondents shared their telephone or online contact details, which were then used to send invitations for the following survey stages. Each invitation included a personal link to the relevant online questionnaires.

3.3. Data collection and data yield

The first wave of the survey took place from October 12 to 16, 2022; the second wave occurred two weeks later, and the third four weeks after the initial survey. To ensure a sufficient sample for retesting, we created a large (oversized) base sample of 2,024 respondents (Test – T), technically representing the RF 18+ population, with a standard error of the initial test sample estimated at 2.12%. Some respondents accepted our offer to participate in follow-up stages and provided their contact details. The second wave (1st retest) gathered 502 observations, and the third wave (2nd retest) 139 observations. While the last two samples

do not represent the general population, they are well-suited for use in experimental tasks. The main socio-demographic characteristics of the final samples are presented in Appendix C. The dropout rate was 75.2% for the first retest and 76.4% for the second. Data suggest some evidence of systematic attrition (Appendix E), with the main factors in participation decline being a lack of interest in the topics and relatively positive attitudes toward "Suicide" and "Domestic Violence" according to "indirect" measures. Self-report variables were neutral toward attrition, while younger respondents were more likely to continue participation compared to older respondents. This specific attrition bias may affect the reliability of GATA tests more than the self-report reliability tests.

The resulting datasets contain only complete observations. "Don't know / No answer / Refused" options were technically disabled. At the start of the self-report task, respondents were instructed to select "the most likely option" when uncertain. In the GATA task, respondents were advised to choose randomly if unsure about their preferred shape. Respondents could opt to end the interview at any point. A total of 2,506 (the test), 510 (the first retest), and 141 (the second retest) respondents started the interview, resulting in 2,024, 502, and 139 completed responses, respectively. Details of unfinished interviews are provided in Appendix D.

Thus, the maximum experimental group consists of the 1st retest with data from 502 respondents' reactions to the same stimuli at a personal level. The minimum experimental group comprises the 2nd retest, corresponding to 139 respondents.

In summary, the effective sample for the first test-retest procedure consists of 502 respondents, and for the second, 139 respondents.

4. The main findings

The general overview of the data is presented in Tables 1 and 2, where the GATA data are displayed alongside similar measurements

from the "direct" self-report questions, which we plan to use as a benchmark against which the results of GATA can be preliminarily assessed.

The values presented in Tables 1 and 2 suggest that the data yield does not show any obvious anomalies. The mean ranges are 2.68 – 7.36 for the "direct" questions and 3.67 – 6.32 for GATA. The group-level consistency of attitudes appears to be similar for both types of measurement, with the standard deviation around 2.0 for both instruments (SD range is 1.003 – 2.665 for the "direct" questions and 1.833 – 2.524 for GATA).

All our variables are ordinal, and none of the distributions are truly normal. All variables failed the Shapiro-Wilk test. Therefore, in order to study the test-retest reliability, we have chosen the criterion of Spearman ρ, which is one of the commonly accepted metrics for ordinal variables [14]. The data collected allow us to conduct 18 reliability tests: three approaches for each of our six variables. They are as follows:

Approach 1: Test vs. 1st Retest. Time distance is two weeks. $N = 502$.

Approach 2: 1st Retest vs. 2nd Retest. Time distance is two weeks. $N = 139$.

Approach 3: Test vs. 2nd Retest. Time distance is four weeks. $N = 139$.

This design means that we have two approaches for assessing short-term reliability (Approaches 1 and 2: two-week interval) and one for assessing long-term reliability (Approach 3: four-week interval).

Statistically, only Approach 1 is based on a sample of a conventionally accepted size. We calculate the coefficients for the two other approaches only as a reference. Due to the inadequacy of the samples, they are not able to provide unquestionable proof for our current hypothesis, but we believe they can help us formulate reasonable assumptions for our further studies.

Table 1

BASIC STATISTICS FOR THE DISTRIBUTIONS OBTAINED FROM THE "DIRECT" QUESTIONS
(8-point ordinal scale; "1" indicates full acceptance of this social practice)

Self – report	Mean			SD		
	first wave ($N = 2024$)	second wave ($N = 502$)	third wave ($N = 139$)	first wave ($N = 2024$)	second wave ($N = 502$)	third wave ($N = 139$)
Suicide	5.52	6.09	5.74	2.596	2.165	2.665
Execution	6.43	6.49	6.18	2.088	1.878	2.007
Taxes avoidance	4.88	3.38	4.73	1.808	1.960	1.852
Corruption	6.74	6.08	6.73	1.760	1.825	1.498
Divorce	2.68	2.90	2.63	1.833	1.936	1.828
Domestic violence	7.36	7.09	7.35	1.003	1.422	0.954

Table 2

BASIC STATISTICS FOR THE DISTRIBUTIONS OBTAINED FROM GATA
(8-point ordinal scale; "1" indicates full acceptance of this social practice)

GATA	Mean			SD		
	first wave ($N = 2024$)	second wave ($N = 502$)	third wave ($N = 139$)	first wave ($N = 2024$)	second wave ($N = 502$)	third wave ($N = 139$)
Suicide	5.78	5.94	5.72	2.105	2.056	2.165
Execution	5.22	5.48	5.10	2.325	2.050	2.341
Taxes avoidance	3.67	4.46	4.73	2.302	2.049	2.229
Corruption	4.66	4.90	4.51	2.524	2.157	2.284
Divorce	6.32	4.29	4.88	1.833	2.122	2.099
Domestic violence	5.17	4.61	4.70	2.078	2.040	2.535

Table 3 presents the Spearman ρ values for all six attitudes measured with "direct" questions, followed by their interpretation as suggested by C. Dancey and J. Reidy [15]. Supporting material for the interpretation of the correlation values is presented in Appendix F. We have applied our own approach to the correlation reference values, but see also appropriate alternatives according to J. Nunnally [16] and D. Hays and colleagues [17].

Table 4 structures the same type of data for GATA measurements.

As the data in Tables 3 and 4 show, the correlations in Approach 1 are generally at almost the same level for both "direct" measurements and GATA. The best ρ value for the self-report is 0.796 ("Suicide"), while for GATA it is 0.762 ("Execution"). The lowest values are 0.190 for self-report ("Tax avoidance") and 0.503 for GATA ("Suicide"). Qualitatively assessing the results, we can find:

"*Very strong*" results: Self-report – 2 ("Suicide," "Domestic violence"); GATA – 2 ("Execution," "Corruption").

"*Strong*": Self-report – 3 ("Execution," "Corruption," "Divorce"); GATA – 4 ("Suicide," "Tax avoidance," "Divorce," "Domestic violence").

"*Moderate*" and "*Weak*": Self-report – 0; GATA – 0.

"*Negligible*": Self-report – 1 ("Tax avoidance"); GATA – 0.

This allows us to conclude that, according to Approach 1 data, GATA and conventional "direct" measurements have a very close level of reliability, which can be generally qualified as acceptable. In terms of attitude objects, this grade is unstable. The GATA output seems to exhibit slightly less variability. It has a relatively lower upper Spearman ρ (0.762 vs. 0.796) and a relatively higher lower Spearman ρ (0.503 vs. 0.190), resulting in a narrower range of the metric: 0.259 vs. 0.606 for self-report.

In practical terms, this means that the reliability of GATA is comparable to that of "direct" measurement. At least for the short period of two weeks or less, it does not generally appear to be significantly better or worse than the self-report method.

Table 3

"DIRECT" QUESTIONS TEST-RETEST RESULTS, SPEARMAN ρ,
C. DANCEY AND J. REIDY INTERPRETATION

Self report	Test vs. 1st Retest ($N = 502$), two weeks distance		1st Retest vs. 2nd Retest ($N = 139$), next two weeks distance		Test vs. 2nd Retest ($N = 139$), four weeks total distance	
	P	Interpretation	ρ	Interpretation	ρ	Interpretation
Suicide	0.796	Very strong	0.649	Strong	0.755	Very strong
Execution	0.573	Strong	0.833	Very strong	0.713	Very strong
Taxes avoidance	0.190	Negligible	0.231	Weak	0.354	Moderate
Corruption	0.465	Strong	0.777	Very strong	0.424	Strong
Divorce	0.620	Strong	0.880	Very strong	0.644	Strong
Domestic violence	0.720	Very strong	0.590	Strong	0.668	Strong

Table 4

GATA TEST – RETEST RESULTS, SPEARMAN ρ,
C. DANCEY AND J. REIDY INTERPRETATION

GATA	Test vs. 1st Retest ($N = 502$), two weeks distance		1st Retest vs. 2nd Retest ($N = 139$), next two weeks distance		Test vs. 2nd Retest ($N = 139$), four weeks total distance	
	P	Interpretation	ρ	Interpretation	ρ	Interpretation
Suicide	0.503	Strong	0.605	Strong	0.245	Weak
Execution	0.762	Very strong	0.882	Very strong	0.778	Very strong
Taxes avoidance	0.692	Strong	0.427	Strong	0.324	Moderate
Corruption	0.741	Very strong	0.401	Strong	0.145	Negligible
Divorce	0.556	Strong	0.778	Very strong	0.320	Moderate
Domestic violence	0.665	Strong	0.160	Negligible	0.267	Weak

Strictly methodologically, we found some evidence of GATA's unsatisfactory reliability. However, according to our best sample from Approach 1, GATA demonstrates two cases of acceptable reliability ("Execution" $\rho = 0.762$ and "Corruption" $\rho = 0.741$). This makes it possible to reject our technical hypothesis (**$H_0$1**): "The test-retest reliability of GATA is less than Spearman $\rho < 0.700$ for every object evaluated." Potentially, GATA is able to demonstrate reliable results.

Let's examine these conclusions using our auxiliary data from Approaches 2 and 3. Conceptually, Approach 2 is comparable to the first, as both represent the same time period of two weeks. In contrast to the set of "direct" measurements, which improves (with 3 "Very strong," 2 "Strong," and 1 "Weak" grades), the GATA output generally remains at the same level: 2 "Very strong" (+0), 3 "Strong" (-1), and 1 "Negligible" (+1) grade. If we look at the measurements of both Approaches 1 and 2 as a single set, we can see that for GATA, 4 out of 12 cases are equal to or exceed our threshold of "Spearman $\rho > 0.700$" for the strength of the test–retest data relationship. This provides additional support for our conclusions regarding (**$H_0$1**).

Approach 3 differs from the others in that the interval between test and retest is longer, in this case, four weeks rather than two. Comparing the data from Approaches 1 and 3 allows us to assess the temporal stability of the "direct" measurement and the reliability of the GATA data. The corresponding data in Table 3 suggest that for the self-report method, the correlations keep their scores slightly apart in absolute terms. In contrast, the GATA correlation drops significantly. Table 4 shows only one attitudinal object that maintains its initial grade ("Execution" – "Very strong," initial value – 0.762, resulting value - 0.778).

This means that in the context of our experiment, the GATA measurements showed a general temporal instability. This led us to reject our (**$H_0$2**): "The test-retest reliability of the GATA retains its initial grade at a time interval of four weeks." All of our data suggest that the results of the GATA retest correlations tend to deteriorate over time. As this occurs while the "direct" measurement correlations remain

relatively stable, this trend should be interpreted as a characteristic of GATA rather than an effect of external factors, such as possible peculiarities of the sample or the attitudinal objects.

Finally, the comparison of all the approaches enables us to evaluate the potential dependence of GATA reliability on the attitudinal objects. Tables 3 and 4 show that both methods have objects with outstanding results. For the self-report, this is "Tax avoidance," which demonstrates atypically weak results for each of the three approaches. For GATA, it is "Execution," which produces atypically strong results that are also stable over time. On the basis of these data, we should conclude that both methods have demonstrated their dependence on the attitudinal objects they seek to evaluate. As far as GATA is concerned, we should reject our (**$H_0$3**): "The test-retest reliability of GATA within the same time interval is of the same grade for each evaluated object."

Thus, all of our technical hypotheses should be rejected. According to that output, the overall substantive conclusions can be presented as follows:

1. To date, GATA is not unquestionably reliable in terms of test-retest reliability. In some cases, it can yield "very strong" results, but in others, it can produce only "weak" or even "negligible" results. The task is therefore to identify possible determinants of this instability. It is reasonable to assume that there are some manageable factors of instability that could potentially be ruled out.

2. In general, GATA reliability tends to deteriorate over time. This may be due to malfunctioning of the measurement procedures or to the natural peculiarities of the GATA measurand. If the second possibility is true, it conflicts with one of the assumptions of GATA, namely that it measures a fraction of attitudes. Within the conventional theoretical framework, it is hard to imagine an "attitude" so unstable as to change every few weeks.

3. GATA is sensitive to the object being evaluated. For some objects, its reliability may be perfect and stable over time, but for others, it may be unpredictably variable. There may be natural limits to

the applicability of GATA. In this case, it might be effective to identify the areas where GATA can be applied with proven reliability and then gradually extend it to still problematic areas.

Therefore, for this stage of GATA validation, we limit ourselves to noting that GATA results show better retest reliability for relatively short periods (up to two weeks) than for longer periods (from four weeks). For these short periods, their retest reliability is comparable to that of the "direct" measures. Overall, we have found no evidence to suggest that the reliability of GATA is fundamentally inadequate.

5. Conclusions

Taking a broad view, one could conclude that our experiment has effectively achieved its basic objectives. We have collected a comprehensive dataset that provides all the means to evaluate GATA in terms of the reliability of its measurements. Taking into consideration the benchmarks set for our sample and attitudinal objects by the "direct" measurements self-report method, GATA showed a comparable level of reproducibility in the short term. Compared to these benchmarks, however, GATA scores relatively low on long-term reliability. This raises questions about GATA's potential for prospective enhancement and development.

Three obvious directions for improving GATA's reliability emerge from our findings:

Random error reduction. As mentioned above, the short-term reliability of GATA is not perfect, but it is quite comparable to that of "direct" measurements. For the latter, this phenomenon has been well studied, and a solution has been found in the construction of a summative scale. A set of relevant and internally reliable variables creates a "Likert space" within which a studied object receives a comprehensive evaluation. The result is an integral summative scale, potentially capable of compensating for contrasting errors in input measures. For GATA, this can be achieved by spreading the dimensions

of the attitudinal object evaluation. Osgood's kit of attitudinal indicators can be used as an instrument for constructing such a set.

Temporal stabilization. The low temporal stability of indirect measurements seems to be a common phenomenon. Some authors suggest accepting it as a natural characteristic of indirect measurements, the negative effects of which can be easily eliminated by averaging the results of several consecutive measurements [18: 6]. In our case, this could be achieved by an additional experiment consisting of a series of GATA measurements. Its results will provide the opportunity to compare the results based on averaging. In practice, however, this approach appears to be of questionable effectiveness, as a typical social study needs to complete a portion of the results for each wave of its fieldwork. Potentially, this may have some theoretical implications in the form of a tentative assumption that the measurand of indirect instruments is some kind of "liquid" fraction of attitudes. Could this particular fraction be related to "true" attitudes? This remains a good question for further discussion.

Adaptation to the target. The reliability of GATA is not the same for different objects that are evaluated. Some of them (such as "Execution" in our experiment) can produce perfect records of reliability level and maintain it over time, while others cannot. Hypothetically, this can be explained by random errors or by the natural characteristics of the objects. The latter hypothesis is supported by one of our previous experiments, in which specific objects (concepts such as "girl," "boy," "man," "grandmother," etc.) were associated with generic concepts such as "men" and "women." It turned out that the sample does not differentiate the elderly with the scale of gender but does so reliably for other objects. This may mean that the typical GATA stimulus apparatus may be inappropriate for certain objects. Improving the method in this area seems possible by expanding the variety of rating dimensions, as mentioned above. This could potentially reduce the effect of probable incongruence between the measuring instrument and the object of evaluation.

All this allows for final methodological and instrumental conclusions. The GATA measurement can potentially produce reliable results. At the moment, there is no evidence that fundamentally compromises this ability for the short period of one or two weeks. In any case, in terms of reliability, the results obtained by GATA over a short period place it on par with the best examples of implicit measures, such as the IAT. Greenwald and Lai's meta-analysis of 58 studies reported that test–retest reliabilities for IAT measures averaged Pearson $r = 0.500$, which can be interpreted as a "strong" correlation according to the De Vaus model. At the same time, this ability is undoubtedly unstable for longer periods. The determinants of this instability are still unclear.

Finally, there is an unplanned observation among our findings that may have some theoretical implications. Namely, the experiment provided further evidence of the orthogonality of GATA and self-report measures. They differ in every characteristic that we compared, from the values of the reliability metric to the dynamics of their changes.

REFERENCES

1. *Fazio, R.* Attitudes as object-evaluation associations of varying strength, Social Cognition, 2007, 25(5), p. 603–637. DOI: 10.1521/soco.2007.25.5.603

2. *Chen, M., Bargh, J.* Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus, Personality and Social Psychology Bulletin, 1999, 25(2), p. 215–224. DOI: 10.1177/014616729902500200

3. L*ikert, R.* Technique for the Measurement of Attitudes, Archives of Psychology, 1932, 140, p. 1–55.

4. *Chernozub, O.* Theory of (Un)Planned Behavior? How our behavioral predictions suffer from "unplanned" actions, The Russian Sociological Review, 2022, 21 (4), p. 82–105. DOI:10.17323/1728-192x-2022-4-82-105.

5. *Gawronski, B., Hahn, A.* Implicit Measures: Procedures, Use, and Interpretation. URL: https://www2.psych.ubc.ca/~schaller/528Readings/GawronskiHahn2019.pdf (date of access: 27.11.2023)

6. *Perugini, M., Richetin, J., Zogmaister, C.* Prediction of behavior. In: Gawronski B., Payne B. (eds.) Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. New York: Guilford Press, 2010. P. 255–277.

7. *De Houwer, J., Moors, A.* How to define and examine the implicitness of implicit measures. In: Wittenbrink B., Schwartz N. Implicit measures of attitudes: Procedures and controversies. Guilford, 2007. P. 179–194.

8. *Gawronski, B., Hahn, A.* Implicit Measures: Procedures, Use, and Interpretation. URL: https://www2.psych.ubc.ca/~schaller/528Readings/GawronskiHahn2019.pdf (date of access: 27.11.2023)

9. *Chernozub, O.* Graphic associative test of attitudes as a convenient implicit measurement tool for mass polls, RUDN Journal of Sociology, 2023, 23 (1), p. 122–141. DOI: 10.22363/2313-2272-2023-23-1-122-141.

10. *Chernozub, O.* Do indirect measures of attitudes improve our predictions of behavior? Evaluating and explaining the predictive validity of GATA, RUDN Journal of Sociology, 2024, 24 (4), p. 241–256. DOI: 10.22363/2313-2272-2024-24-1-241-258.

11. *Chernozub, O.* Affective components of electoral behavior: design and validity of visual association test of attitude (in Russian), Monitoring of Public Opinion: Economic and Social Changes, 2018, 3, p. 3–28. DOI: 10.14515/monitoring.2018.3.01.

12. *Chernozub, O.* The two-component model of behavior factors: evidences of orthogonality of explicit and implicit factors, RUDN Journal of Sociology, 2022, 22 (1), p. 70–83. DOI: 10.22363/2313-2272-2022-22-1-70-83.

13. *Cicchetti, D.* Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology, Psychological Assessment, 1994, 6 (4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

14. *Kruskal, W.* Ordinal Measures of Association, Journal of the American Statistical Association, 1958, 53 (284), p. 814–861. DOI:10.2307/2281954.

15. *Dancey, C., Reidy, J.* Statistics Without Maths for Psychology. Pearson, 2020.

16. *Nunnally, J.* Psychometric Theory, 2nd ed. New York: McGraw-Hill, 1978.

17. *Hays, R.D., Anderson, R., Revicki, D.* Psychometric considerations in evaluating health-related quality of life measures, Quality of Life Research, 1993, 2 (6), p. 441–449. DOI:10.1007/BF00422218

18. *Moors, A., Koster, M.* Behavior prediction requires implicit measures of stimulus-goal discrepancies and expected utilities of behavior options rather than of attitudes toward objects, Wiley interdisciplinary reviews. Cognitive science, 2022, 13 (5). DOI: 10.1002/wcs.1611

Information about the authors

Oleg L. Chernozub

Ph.D. in Sociology; Lead Researcher, Institute of Sociology FCTAS RAS, Moscow, Russia

ResearcherID: GNH-5045-2022

Appendix

Appendix A. INSTRUMENTS OF INDIRECT MEASUREMENTS

1. Name Letter Task (Nuttin, 1985, 1987; see Lebel et al., 2009).
2. Evaluative Priming Task (Fazio et al., 1986).
3. Linguistic Intergroup Bias (Maass, Salvi, Arcuri, & Semin, 1989).
4. Implicit Association Test (Greenwald, McGhee, & Schwarz, 1998).
5. Approach-Avoidance Tasks (e.g., Chen & Bargh, 1999; Castelli, Zogmaister, Smith, & Arcuri, 2004).
6. Go/No-Go Association Task (Nosek & Banaji, 2001).
7. Weapon Paradigm (Payne, 2001; Correll, Park, Judd, & Wittenbrink, 2002).
8. Extrinsic Affective Simon Task (De Houwer, 2003).
9. Pesonalized IAT (Olson & Fazio, 2004).
10. Affect Misattribution Procedure (Payne et al., 2005).
11. Evaluative Movement Assessment (Brendl et al., 2005).
12. Implicit Association Procedure (Schnabel et al., 2006).
13. Single Category IAT (Karpinski & Hilton, 2006).
14. Identification Extrinsic Affective Simon Task (De Houwer & De Bruycker, 2007).
15. Single Block IAT (Teige-Mocigemba et al, 2008).
16. Brief IAT (Sriram & Greenwald, 2009).
17. Recoding Free IAT (Rothermund et al., 2009).
18. Sorting Paired Features Task (Bar-Anan et al., 2009).
19. Action Interference Paradigm (Banse et al, 2010).
20. Implicit Relational Assessment Procedure (BarnesHolmes et al., 2010).

Appendix B. WORDING OF THE STIMULI

Basic instructions:

For self – report ("direct" – measurement) variables: "Please make a note for each of the following actions whether you think it can always be justified, never be justified, or something in between. (On an 8-point scale, 1 – never justifiable; 8 – always justifiable)".

For GATA ("indirect" measurement) variables: "Please read the word and chose the graphical shape, which is the most suitable" / "Please rate the

shapes starting with those that you would like to look at, have around or touch and ending with those that would be unpleasant to look at, unpleasant to have around or unpleasant to touch". (This results in an 8-point scale; 1 – the least implicitly preferred object, 8 – the most implicitly preferred object).

Actions for assessment / "Words" for association with graphical shapes.

1. Suicide.
2. Execution.
3. Tax avoidance.
4. Corruption.
5. Divorce.
6. Domestic violence.

Appendix C. SOCIODEMOGRAPHIC CHARACTERISTICS OF THE SAMPLES

Table C.1

SOCIODEMOGRAPHIC CHARACTERISTICS, CONTROLLED WITHIN THE EXPERIMENT

	T	R1	R2
N	2024	503	139
Sex	100.0%	100.0%	100.0%
Male	51.7%	60.6%	62.8%
Female	48.3%	39.4%	37.2%
Age	100.0%	100.0%	100.0%
18–24	5.6%	4.0%	5.3%
25–34	13.5%	25.9%	31.9%
35–44	17.6%	27.1%	28.7%
45–54	17.9%	17.1%	12.8%
55–64	22.3%	17.9%	13.3%
65+	23.1%	8.0%	8.0%
Occupation	100.0%	100.0%	100.0%
Employed	58.8%	60.6%	63.3%
Student	2.6%	4.0%	2.7%
Unemployed	6.4%	9.6%	12.8%

End of tab. C.1

N	T 2024	R1 503	R2 139
Houskeeper	6.2%	8.0%	8.0%
Retired	21.8%	15.9%	10.6%
Other	4.2%	2.0%	2.7%

Appendix D. COMPLETION RATES

Table D.1

COMPLETION RATES FOR ONLINE INTERVIEW
BY STAGES OF INTERRUPTING

Initial samples	T 2506	R1 510	R2 141
Respondents			
SR	64	2	0
GATA – associaton	112	2	0
GATA – ranking	32	1	0
GATA – total	144	3	0
Other	274	2	2
Total incomplits	482	7	2
% of the initial sample			
SR	2.6%	0.4%	0.0%
GATA – associaton	4.5%	0.4%	0.0%
GATA – ranking	1.3%	0.2%	0.0%
GATA – total	5.7%	0.6%	0.0%
Other	10.9%	0.4%	1.4%
Total incomplits	19.2%	1.4%	1.4%
Effective samples	2024	503	139

Appendix E. ATTRITION

Table E.1

PERTICIPATION STATUS (SPLIT OFF OR STAYED) VS. MAIN VARIABLES. STATUS AS INDEPENDENT VARIABLE, SOMERS *D* VALUE, ρ.

	Test vs. 1st Retest		1st Retest vs 2nd Retest		Who are more likely to stay ($\rho < 0.05$)
	Somers *D*	ρ	Somers *D*	ρ	
Socio-demographic					
Age	0.298	0.000	0.306	0.001	Young resp.
Gender	-0.007	0.903	0.028	0.742	
Interest towards the issue					
Suicide	0.467	0.000	0.517	0.000	High interest
Execution	0.100	0.015	0.096	0.098	High interest
Taxes avoidance	0.275	0.000	0.311	0.000	High interest
Corruption	0.134	0.046	0.184	0.050	High interest
Divorce	0.002	0.964	0.036	0.663	
Domestic violence	0.186	0.000	0.188	0.013	High interest
Self report					
Suicide	0.044	0.447	0.102	0.241	
Execution	0.133	0.085	0.186	0.100	
Taxes avoidance	-0.043	0.508	-0.026	0.762	
Corruption	-0.086	0.196	-0.061	0.518	
Divorce	0.139	0.092	0.167	0.065	
Domestic violence	0.121	0.249	-0.085	0.357	
GATA					
Suicide	0.185	0.001	0.234	0.006	Negative att.
Execution	0.130	0.090	0.157	0.053	
Taxes avoidance	0.084	0.197	0.081	0.377	
Corruption	0.105	0.120	0.158	0.105	

End of tab. E.1

	Test vs. 1st Retest		1st Retest vs 2nd Retest		Who are more likely to stay ($\rho < 0.05$)
	Somers *D*	ρ	Somers *D*	ρ	
Divorce	0.086	0.278	0.091	0.424	
Domestic violence	0.396	0.000	0.474	0.000	Negative att.

Appendix F. CORRELATION INTERPRETATION

Table F.1

CORRELATION INTERPRETATION BY DE VAUS

Pearson *r*	Correlation Strength
0.00	No Correletion
0.01–0.09	Non-significant Correlation
0.10–0.29	Weak Correlation
0.30–0.49	Moderate Correlation
0.50–0.69	Strong Correlation
0.70–0.89	Very Strong Correlation
> 0.9	Almost Perfect Correlation

Adopted: *De Vaus D.* Surveys in Social Research. London: Routledge, 2002. 422 p.

Table F.2

CORRELATION INTERPRETATION BY DANCEY AND REIDY

Spearman ρ	Correlation
0.01–0.19	No or negligible relationship
0.2–0.29	Weak relationship
0.3–0.39	Moderate relationship
0.4–0.69	Strong relationship
≥ 0.70	Very strong relationship

Adopted: *Dancey C., Reidy J.* Statistics Without Maths for Psychology. Pearson 2020. 640 p.

DOI: 10.19181/4m.2023.32.2.4

ГАТО: ПРОВЕРКА НАДЕЖНОСТИ ИЗМЕРЕНИЙ МЕТОДОМ ПОВТОРНОГО ТЕСТИРОВАНИЯ

Чернозуб Олег Леонидович
Институт социологии ФНИСЦ РАН, Москва, Россия
9166908616@mail.ru
ORCID: 0000-0001-5689-8719

Для цитирования: *Чернозуб О. Л.* ГАТО: проверка надежности измерений методом повторного тестирования // Социология: методология, методы, математическое моделирование (Социология:4М). 2023. № 57. С. 112–137. DOI: 10.19181/4m.2023.32.2.4. EDN: PWCWQK.

Аннотация. Социальные исследователи долгое время стремились преодолеть уязвимость метода самоотчета к различным эффектам, основанным на неспособности или нежелании респондентов отвечать правильно. Одна из очевидных стратегий решения этой проблемы состоит в том, чтобы попытаться извлечь информацию таким образом, чтобы оценить отношение респондента к исследуемому объекту по его действиям, «косвенно» связанным с этим объектом, не опираясь на «прямые» ответы респондента о том, каково, по его мнению, это отношение. Одним из многочисленных инструментов, которые, претендуют на способность исследовать установки респондента подобным образом, является «Графический ассоциативный тест отношения» (ГАТО).

В этой статье представлены основные итоги анализа ретестовой надежности ГАТО. 18 повторных тестов через две (12 тестов) и четыре (6 тестов) недели после первоначального измерения показали, что ГАТО потенциально может давать надежные результаты. В то же время, эта способность не вполне стабильна, выявлена зависимость от объекта оценивания и периода времени между сравниваемыми тестами. Если для двухнедельных периодов надежность ГАТО довольно часто демонстрирует удовлетворительный уровень, (Спирмен $\rho > 0{,}700$ в 4 случаев из 12) и сопоставима с методом самоотчета, то для четырехнедельного периода оценки только 1 объекта из 6 сохраняют настолько же высокий уровень надёжности.

В статье также обсуждаются возможные направления будущих исследований, направленных на увеличение надёжности измерений ГАТО.
Ключевые слова: метод самоотчета; прямые измерения; косвенные измерения; тест – ретестовая надежность; ГАТО

Литература

1. *Fazio R.* Attitudes as object-evaluation associations of varying strength // Social Cognition. 2007. Vol. 25, No. 5. P. 603–637. DOI: 10.1521/soco.2007.25.5.603
2. *Chen M., Bargh J.* Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus // Personality and Social Psychology Bulletin. 1999. Vol. 25, No. 2. P. 215–224. DOI: 10.1177/0146167299025002000. EDN: JPAUBJ.
3. *Likert R.* Technique for the Measurement of Attitudes // Archives of Psychology. 1932. Vol. 140. P. 1–55.
4. *Chernozub O.* Theory of (Un)Planned Behavior? How our behavioral predictions suffer from “unplanned” actions // The Russian Sociological Review. 2022. Vol. 21, No. 4. P. 82–105. DOI: 10.17323/1728-192x-2022-4-82-105. EDN: HIAPKI.
5. *Gawronski B., Hahn A.* Implicit Measures: Procedures, Use, and Interpretation. URL: https://www2.psych.ubc.ca/~schaller/528Readings/GawronskiHahn2019.pdf (date of access: 27.11.2023).
6. *Perugini M., Richetin J., Zogmaister C.* Prediction of behavior // Handbook of Implicit Social Cognition: Measurement, Theory, and Applications / Ed. Gawronski B., Payne B. New York: Guilford Press, 2010. P. 255–277.
7. *De Houwer J., Moors A.* How to define and examine the implicitness of implicit measures // Implicit measures of attitudes: Procedures and controversies / Ed. Wittenbrink B., Schwartz N. Guilford, 2007. P. 179–194.
8. *Gawronski B., Hahn A.* Implicit Measures: Procedures, Use, and Interpretation. URL: https://www2.psych.ubc.ca/~schaller/528Readings/GawronskiHahn2019.pdf (date of access: 27.11.2023).
9. *Chernozub O.* Graphic associative test of attitudes as a convenient implicit measurement tool for mass polls // Вестник РУДН. Серия: Социология. 2023. Т. 23, № 1. С. 122–141. DOI: 10.22363/2313-2272-2023-23-1-122-141. EDN: QPWGNB.

10. *Chernozub O.* Do indirect measures of attitudes improve our predictions of behavior? Evaluating and explaining the predictive validity of GATA // Вестник РУДН. Серия: Социология. 2024. Т. 24, № 4. С. 241–256. DOI: 10.22363/2313-2272-2024-24-1-241-258. EDN: ZTTKWR
11. *Чернозуб О.* Л. Выявление аффективной компоненты электоральной установки: создание и валидизация графического ассоциативного теста отношения // Мониторинг общественного мнения: Экономические и социальные перемены. 2018. № 3. С. 3–28. 10.14515/monitoring.2018.3.01. EDN: XSWAPR.
12. *Chernozub O.* The two-component model of behavior factors: evidences of orthogonality of explicit and implicit factors // Вестник РУДН. Серия: Социология. 2022. Т. 22, № 1. С. 70–83. DOI: 10.22363/2313-2272-2022-22-1-70-83. EDN: PCHJEF.
13. *Cicchetti D.* Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology // Psychological Assessment. 1994. Vol. 6, No. 4. P. 284–290. DOI: 10.1037/1040-3590.6.4.284
14. *Kruskal W.* Ordinal Measures of Association // Journal of the American Statistical Association. 1958. Vol. 53, No. 284. P. 814–861. DOI: 10.2307/2281954.
15. *Dancey C., Reidy J.* Statistics Without Maths for Psychology. Pearson 2020.
16. *Nunnally J.* Psychometric Theory, 2nd ed. New York: McGraw-Hill, 1978.
17. *Hays R.D., Anderson R., Revicki D.* Psychometric considerations in evaluating health-related quality of life measures // Quality of Life Research. 1993. Vol. 2, No. 6. P. 441–449. DOI: 10.1007/BF00422218. EDN: QVQEJF.
18. *Moors A., Koster M.* Behavior prediction requires implicit measures of stimulus-goal discrepancies and expected utilities of behavior options rather than of attitudes toward objects // Wiley interdisciplinary reviews. Cognitive science. 2022. Vol. 13, No. 5. DOI: 10.1002/wcs.1611. EDN: EMJRRJ.

Информация об авторах

Олег Леонидович Чернозуб
кандидат социологических наук,
ведущий научный сотрудник Института социологии ФНИСЦ РАН
ResearcherID: GNH-5045-2022