METHODOLOGY | DATA QUALITY | BEHAVIORAL RESEARCH

A Modular Data Quality Framework for Online Behavioral Research

Design, Implementation, and Application in Web-Based Implicit Association Testing

AT A GLANCE

The Headway Data Quality Framework is a modular, five-component system operationalizing research-grade quality assurance for online behavioral research. Deployed in the KMOP Gender-Career IAT (Greece, N = 1,145), it demonstrates practical value across the full data lifecycle - from session validation through precision-informed recruitment strategy. Available as a deployable research infrastructure service.

AUTHOR

Yiannis Pappas

Headway S.A.

CASE STUDY

KMOP Gender-Career IAT

Greece, 2025-2026

DOI

https://doi.org/10.5281/zenodo.19358631

Abstract

Online behavioral research is expanding rapidly across applied settings. NGOs, policy institutes, and research SMEs now routinely commission experience sampling instruments, implicit association studies, and reaction-time tasks that were until recently confined to academic laboratories. Yet the infrastructure for ensuring data quality in these deployments has not kept pace. Most organizations collecting behavioral data online lack documented quality assurance pipelines, apply inconsistent exclusion criteria, and present subgroup findings without reference to statistical adequacy - producing results that are difficult to defend under scrutiny and ill-suited to open science disclosure requirements increasingly expected by funders and clients.

This paper presents the Headway Data Quality Framework: a modular, five-component system that Headway S.A. (hereafter Headway) developed and deployed to address this gap for organizations conducting Implicit Association Tests (IATs) and analogous online behavioral instruments. The framework operationalizes quality assurance across the full research lifecycle - from raw session validation through temporal stability monitoring - transforming quality control from a post-hoc cleanup activity into an active input to research design and stakeholder communication.

We document the framework's deployment in the KMOP Gender-Career IAT, a large-scale civic study examining implicit gender-career associations among Greek adults (N > 1,100 at Stage 1). The deployment demonstrates the framework's practical value: Flagging quality issues at the session level, confirming the analytical comparability of pilot and public-launch data, classifying subgroup analyses by statistical adequacy, monitoring for sample drift across a ten-week open recruitment window, and generating a precision-informed recruitment strategy for Stage 2 confirmatory analysis.

The Headway framework addresses a gap explicitly identified in the methodological literature: Web-based, individual-completed data collection practice is fragmented, inconsistent, and lacks the integrated quality assurance infrastructure that other behavioral research traditions have developed. For collaborative research project consortium partners requiring rigorous, transparent, and funder-defensible behavioral data collection, Headway's framework provides that infrastructure as a deployable service.

Transparency Note

Throughout this paper, 'we' refers to the Headway research and development team.

This paper was developed with the assistance of Claude, an AI assistant developed by Anthropic, for drafting and structural writing. The framework, methodological decisions, analytical content, and all proprietary concepts are the intellectual work of the Headway team. All AI-assisted text was reviewed and validated by the author.

1. Introduction

The behavioral and social sciences are in the middle of a quiet infrastructure crisis. Research instruments that once required controlled laboratory conditions - including reaction-time tasks, implicit association measures, and experience sampling protocols - are now routinely deployed online, reaching thousands of participants through web platforms, organizational networks, and media-driven recruitment campaigns. The scale this enables is real and valuable. So is the problem it creates.

Crowdsourced and open-recruitment online behavioral studies operate without experimenter oversight, on heterogeneous devices, with participants of unknown prior experience and variable motivation. The evidence base for why this matters is substantial and consistent. Connors, Spangenberg, Perkins, and Forehand (2020) showed directly that IATs collected through crowdsourcing platforms produce distorted reliability estimates and weaker criterion validity compared to controlled settings, driven by higher rates of inattention, non-naivety (prior exposure to similar methodologies), and low task motivation. Carpenter et al. (2019) validated a JavaScript/Qualtrics survey-software IAT against IATs run via dedicated reaction-time software (Inquisit), demonstrating adequate psychometric equivalence and producing nearly identical D-scores. Richetin and colleagues (2015) systematically tested 420 scoring parameter combinations on large online datasets and found substantial variability in resulting D-scores.

The consequence is not merely academic. Organizations that commission behavioral research - NGOs building advocacy campaigns, policy institutes evaluating interventions, research SMEs delivering Horizon Europe deliverables - depend on findings that can survive scrutiny, communicate clearly to non-specialist audiences, and meet the open science disclosure standards now expected by major funders. When quality assurance is ad hoc, undocumented, or absent, those conditions are difficult to meet. The problem is widely enough recognized that published best-practice guidelines and checklists for online behavioral research (Gagné & Franzen, 2023; Rodd, 2024; Kochari, 2019) have begun to converge on multi-layer recommendations - recommendations that, in our view, point toward the kind of standardized reporting infrastructure that CONSORT provides for clinical trials. What does not exist is an integrated framework that assembles these best practices into a systematic quality assurance workflow.

Headway built one.

This paper presents the Headway Data Quality Framework: a modular, five-component system designed to operationalize research-grade quality assurance for online IAT research and, by design principle, for online behavioral instruments more broadly. The framework transforms quality assurance from a post-hoc cleanup activity into an integrated feature of research design - one that changes what a research team can claim, document, and defend about its findings.

We document the framework through its application in the KMOP Gender-Career IAT, a large-scale civic study examining implicit associations between gender and career versus family domains among Greek adults, commissioned by KMOP - a research-active non-governmental organization with more than 45 years of experience in social inclusion, gender equality, and labor market policy. The collaboration exemplifies a partnership model of growing relevance in

European research: A domain expert with community infrastructure and communication reach, combined with a methodological specialist with the technical and analytical capability to deliver publication-ready behavioral data. The Headway framework was not bolted onto this project after the fact. It was the infrastructure on which the research ran.

2. Background: Why Online IAT Quality Assurance Is Hard, and Why Existing Practice Falls Short

2.1. The IAT as a Research Instrument at Scale

The Implicit Association Test (Greenwald et al., 1998) measures automatic cognitive associations via response latency contrasts, operationalized as a standardized mean difference (the D-score; Greenwald et al., 2003). Its core logic is elegant: If two concepts are strongly associated, pairing them in a response task produces faster reaction times than pairing concepts that are weakly associated. The D-score algorithm (Greenwald, Nosek, & Banaji, 2003) operationalizes this contrast as a standardized mean difference across response blocks, providing a participant-level estimate of implicit association strength.

The development of web-based IAT platforms transformed the instrument from a laboratory tool into a mass-participation research resource. Project Implicit (Nosek, Banaji, & Greenwald, 2002) demonstrated that IATs could be delivered at mass scale via web browser - accumulating 600,000 completed tests in the first 19 months of operation - without compromising the basic psychometric structure of the instrument. That demonstration has since been replicated extensively: Meta-analytic evidence broadly supports the IAT's capacity to predict intergroup behavior, though effect sizes are modest and heterogeneous (Kurdi et al., 2019), and best-practice consensus documents have been developed by the instrument's originators (Greenwald et al., 2022).

However, the transition to web-based administration at scale introduced data quality challenges that laboratory research largely sidesteps. In the laboratory, an experimenter is present. Equipment is controlled. Participants are typically recruited through structured channels, are naive to the instrument, and complete the task in a defined context. None of these conditions reliably hold in open-recruitment web studies.

2.2.What the Literature Shows About Web-Based IAT Quality

The methodological literature on web-based IATs has accumulated around several well-documented problems, each with practical implications for any organization running an online IAT study.

Scoring variability. Richetin and colleagues (2015) tested 420 distinct combinations of scoring parameters on large online datasets and found substantial variability in resulting D-scores

depending on how extreme latencies were handled, whether error trials were retained, and which trial segments were included in the computation. Their recommendations - replace rather than delete extreme latencies, retain error information with a penalty, use the D-score family of algorithms - are not universally applied in practice. An organization that does not know which scoring choices it is making cannot know how comparable its estimates are to published benchmarks.

Implementation and platform variability. Carpenter et al. (2019) validated a JavaScript-based web IAT against laboratory implementations, demonstrating adequate psychometric equivalence and concluding that survey-software IATs can produce results nearly identical to dedicated platforms, though they noted that implementation details should be explicitly reported. More broadly, Peer et al. (2022) showed that participant source is a major determinant of data quality: Data from Prolific showed consistently high quality on attention, comprehension, and reliability measures; MTurk data showed meaningfully poorer quality even after standard filtering. Organizations recruiting through open media campaigns - with no platform-side vetting whatsoever - face the most methodologically challenging scenario the literature describes.

Participant-level quality. Connors, Spangenberg, Perkins, and Forehand (2020) quantified what poor-quality participation actually does to IAT data: Inattentive responding, prior IAT exposure, and low task motivation each independently distorted reliability and criterion validity in crowdsourced samples. Standard Greenwald exclusion criteria, when applied to MTurk samples, identify dramatically higher proportions of low-quality sessions than in laboratory settings (23–37% vs. 8.9% benchmark; Connors et al., 2020), reflecting genuine differences in sample composition rather than insufficiency of the criteria themselves.

Temporal dynamics. Nosek et al. (2002) observed in early Project Implicit data that media coverage generated substantial recruitment spikes - over 150,000 visits in the five days following two consecutive televised broadcasts - raising concerns (in our reading of their data) about potential shifts in sample composition across recruitment channels. More recent work by Chmielewski and Kucker (2019) documented that participant pool characteristics on MTurk shifted substantially across a multi-wave observation period, suggesting that temporal monitoring may be valuable for studies with extended collection windows.

2.3.Why Existing Practice Falls Short

The literature has not been silent on these challenges. Guidelines for online cognitive and reaction-time research (Gagné & Franzen, 2023; Rodd, 2024; Kochari, 2019) converge on a recommended multi-layer approach: Pre-task device and platform controls, within-task attention management and timing diagnostics, and post-task exclusion pipelines with documented criteria. Modern web browsers achieve timing precision adequate for IAT administration under favorable conditions, though with measurable variability across OS/browser combinations (Bridges et al., 2020; Anwyl-Irvine et al., 2020; Reimers & Stewart, 2015).

What does not exist, as the literature itself acknowledges, is an integrated operational framework that assembles these best practices into a documented, deployable quality

assurance pipeline. Published studies apply individual elements inconsistently. Organizations without in-house methodological expertise have no practical reference for what a complete quality assurance system looks like. And the field lacks a formal equivalent of CONSORT or APA JARS for online reaction-time and IAT research - the kind of reporting standard that makes quality assurance checkable and comparable across studies.

That is the gap Headway's framework addresses.

3. The Headway Data Quality Framework

The Headway Data Quality Framework was built in response to a specific problem: Organizations running online studies in applied settings had no integrated system for managing the quality challenges the literature documents. Individual best practices existed in scattered form across the academic literature. What did not exist was an operational pipeline that assembled those practices into a single documented system, implemented it in real research infrastructure, and produced outputs that research teams could act on in real time.

Headway's framework operationalizes what the literature recommends but left unimplemented. Each of its five components addresses a distinct, documented quality challenge; together they cover the full data lifecycle from raw session ingestion to recruitment strategy. The components are modular - organizations may deploy individual components where specific needs are acute - but they are designed to function as an integrated pipeline when deployed together. Each component produces a documented output that feeds directly into a research decision: flag → review → exclude, segment, or apply sensitivity analysis. Quality assurance, in this framework, is not a cleanup step, but rather a decision architecture.

The framework was developed and is currently implemented in the context of IAT research, which serves here as an illustrative implementation of its broader logic. The design principles underlying each component are deliberately instrument-agnostic: Validity gating, phase comparability checking, precision-tiered subgroup reporting, temporal stability monitoring, and recruitment gap analysis apply to any behavioral instrument collected in an open online setting. The current implementation language (response latency thresholds, D-score computations) is IAT-specific; the logic generalizes.

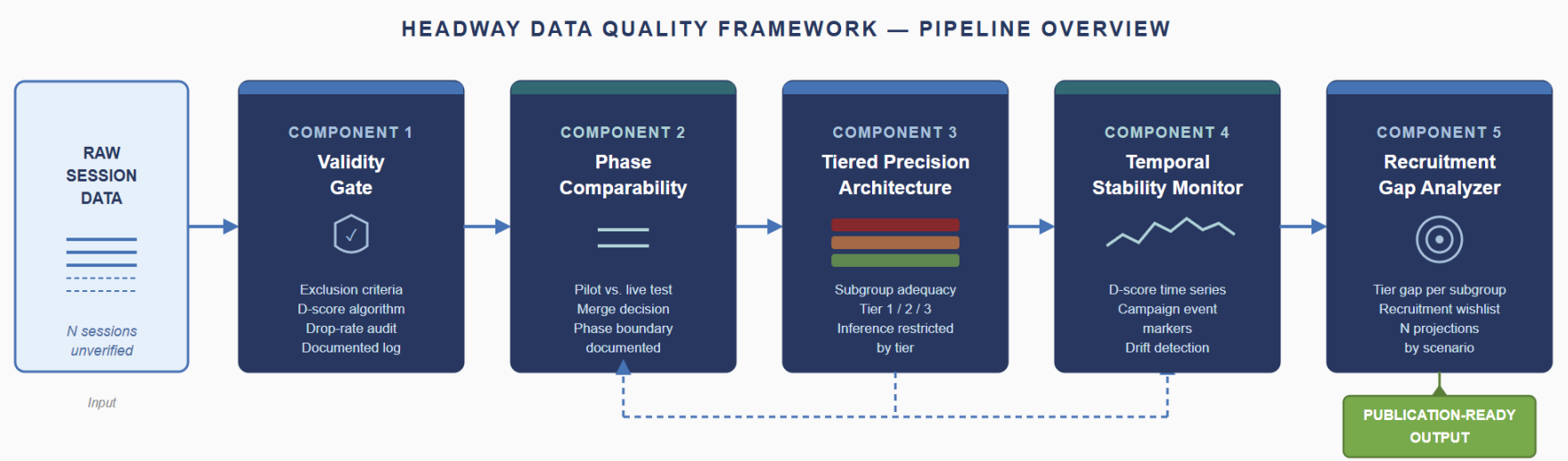


Figure 1. Headway Data Quality Framework — Pipeline Overview.

Component	Function	What it addresses	Literature anchor
1. Validity Gate	Applies IAT exclusion criteria systematically with documented decision logging	Error trials, extreme outliers, fast-responder artifacts	Greenwald et al. (2003); Connors et al. (2020)
2. Phase Comparability Check	Tests statistically whether pilot data can be merged with live data	Sample confounds introduced at deployment phase boundaries	Nosek et al. (2002); Bartmess & Abreu (2024)
3. Tiered Precision Architecture	Classifies subgroups into adequacy tiers; restricts inference by tier	Over-interpretation of small-N subgroup estimates	Cohen (1988); Greenwald et al. (2003); Nosek et al. (2002)
4. Temporal Stability Monitor	Tracks the primary outcome measure as a time series across the recruitment window	Sample drift from media events, targeted recruitment, evolving participant pools	Nosek et al. (2002); Chmielewski & Kucker (2019) - operationalized by Headway
5. Recruitment Gap Analyzer	Identifies subgroups closest to the next tier threshold; produces targeted recruitment wishlist	Post-hoc subgroup adequacy gaps that cannot be recovered after collection	Peer et al. (2022); Uittenhove et al. (2023)

Component 1: Validity Gate

The Validity Gate operationalizes session exclusion and data cleaning using pre-specified, documented criteria applied systematically to every collected session. For IAT research, this implements the Greenwald et al. (2003) improved scoring algorithm - removing trials exceeding 10,000ms, flagging participants with more than 10% of trials below 300ms - while extending beyond the minimum standard to address the quality challenges documented in applied online settings.

Richetin et al. (2015) demonstrated empirically that extreme-latency handling and error-trial treatment substantially affect D-score estimates - supporting the framework's commitment to documenting these choices. Carpenter et al. (2019) validated survey-software IATs against another reaction-time software platform, using the standard Greenwald RT cutoffs (>10% trials <300ms drop) - illustrating that systematic application of documented exclusion criteria is feasible across delivery modalities.

Beyond latency-based criteria, the Gate implements automated drop-out detection: flagging sessions with response patterns inconsistent with engaged participation catches low-quality sessions that latency criteria alone miss. The Gate therefore operates on two layers simultaneously - statistical exclusion criteria and behavioral engagement criteria - with each decision logged to a reproducible audit trail.

The design principle is reproducibility. **Every exclusion decision is governed by pre-specified, documented criteria, enabling any analyst to reconstruct the clean dataset from raw data**. This is a minimum requirement for open science compliance and is increasingly expected by Horizon Europe funders and publishing outlets alike.

Component 2: Phase Comparability Check

Many research deployments proceed through identifiable phases: A pilot or soft-launch period during which the instrument is tested with a limited audience, followed by a full public launch. The methodological question this creates “can data from these phases be merged without introducing systematic bias?” is rarely asked explicitly, yet the consequences of ignoring it are real.

Participants who encounter a study during its soft launch typically arrive through researcher networks, professional contacts, and early-adopter channels. They differ from participants who arrive after a public media campaign in ways that can affect the primary outcome measure. This is not a speculative concern: Chmielewski & Kucker (2019) documented that participant pool characteristics on MTurk shifted substantially over a multi-year observation window, with measurable effects on data quality.

The Phase Comparability Check applies formal tests - mean D-score comparison, variance comparison, demographic profile comparison - **across identified recruitment phases before merging**. Where systematic differences are detected, the component produces a

documentation record and a recommendation: either segment the analysis by phase, apply statistical controls for phase membership, or restrict the primary analysis to the post-launch dataset. The decision is transparent, reproducible, and defensible under methodological scrutiny.

Component 3: Tiered Precision Architecture

The Tiered Precision Architecture addresses what is perhaps the most common source of over-interpretation in applied IAT research: Subgroup estimates presented without reference to the statistical precision those estimates carry at the observed sample size. A finding described as showing "a significant gender difference in implicit career associations" means something very different when it rests on 800 participants per group versus 40.

The architecture classifies each planned subgroup analysis into one of five adequacy bands (two sub-threshold and three named tiers) based on the adequacy of the available sample for the inferential claim being made. Tier 0 (Privacy, N<20) prohibits reporting of subgroup means entirely, consistent with GDPR guidelines for aggregate data anonymisation at very small cell sizes. Between N = 20 and the Tier 1 threshold (Unstable), subgroup estimates are considered too imprecise for substantive interpretation and are excluded from published findings. Tier 1 (Descriptive) permits the reporting of point estimates with uncertainty quantification - confidence intervals, bootstrapped distributions - but precludes formal comparative claims. Tier 2 (Comparative) permits between-group statistical comparisons with adequate power for medium-to-large effects. Tier 3 (Benchmark) supports robust multi-predictor modeling and fine-grained subgroup decompositions.

Tier thresholds are calibrated to the specific variance characteristics of the primary outcome measure and the effect sizes of theoretical and practical interest - a calibration that requires instrument-specific knowledge but follows a general logic applicable to any behavioral measure.

The practical implication: A research team working with a growing dataset can report confidently on high-adequacy subgroups while being transparent about which comparisons require additional data - exactly the communication task that Stage 1 / Stage 2 open science designs require.

Component 4: Temporal Stability Monitor

Web-based studies with open recruitment windows are subject to a form of sample composition drift that is rarely monitored: The population of participants changes as recruitment channels change, and this change can shift the estimated population parameter in ways that may not be immediately visible in cross-sectional results. Nosek and colleagues (2002) observed this dynamic in early Project Implicit data; Chmielewski and Kucker (2019) documented analogous dynamics in MTurk data quality over a multi-year observation window.

The concern is well-established in the literature. What did not exist, prior to Headway's framework, was an operational monitoring system that tracks it in real time. The Temporal Stability Monitor is Headway's answer to that gap.

The component tracks the rolling mean D-score across the full recruitment period, computed over configurable time windows, against control limits derived from the overall distribution. It produces a time-series visualization of estimate stability and flags windows where aggregate values shift beyond expected sampling variation. The output supports two functions: Retrospective interpretation (did a specific recruitment event introduce a compositionally distinct participant wave?) and prospective quality management (is the current estimate stable enough to treat the accumulating dataset as a single analytical unit?).

This component represents the framework's strongest contribution: It addresses a methodological concern that prior literature had identified but not operationalized as a routine data-collection workflow tool.

Component 5: Recruitment Gap Analyzer

The Recruitment Gap Analyzer transforms quality monitoring from a retrospective activity into a prospective research strategy. **Drawing on the Tiered Precision Architecture, it identifies which analytical subgroups are closest to their next adequacy tier threshold and calculates the number of additional participants - disaggregated by demographic profile - required to cross it**.

The component produces a prioritized gap map: A ranked list of subgroups by recruitment priority, expressed in concrete terms (e.g., "34 additional male participants aged 45–54 in private-sector employment would move the age-by-sector interaction analysis from Tier 1 to Tier 2"). This output has direct operational value: it translates statistical needs into recruitment targets that communications and outreach teams can act on.

The component also produces growth projections under different total-N scenarios, enabling the research team to anticipate which analyses will be feasible at different recruitment milestones and to preregister conditional analysis plans accordingly. Uittenhove et al. (2023) and Peer et al. (2022) have both shown that participant source affects sample composition in ways that matter for subgroup adequacy - the Recruitment Gap Analyzer makes those composition gaps visible and actionable before the collection window closes.

4. Case Application: The KMOP Gender-Career IAT

4.1.Study Context

The KMOP Gender-Career IAT is a large-scale civic study examining implicit associations between gender (male/female) and career versus family domains among the Greek adult population. KMOP - a research-active NGO with more than 45 years of experience in social inclusion, gender equality, and labor market policy - commissioned Headway to design, build, and deploy the IAT platform and to provide full analytical quality assurance for the resulting dataset.

The study adopts a two-stage open science design. Stage 1 constitutes an exploratory analysis conducted on the initial accumulating dataset (N > 1,100 at the time of this report), with findings reported with appropriate uncertainty quantification and precision tiering. Stage 2 constitutes a preregistered confirmatory analysis on the full dataset following a media-driven recruitment phase expected to bring total N above 2,000. The Stage 2 preregistration is publicly available on the Open Science Framework, specifying in advance the hypotheses, analysis plan, and exclusion criteria that will govern the confirmatory results.

The KMOP-Headway collaboration instantiates a partnership model of growing relevance in European research funding: a substantive domain expert with community infrastructure and communication reach, paired with a methodological specialist providing research-grade technical and analytical capability. Headway's framework was not an add-on to this project. It was the infrastructure on which the research ran - influencing instrument design, data collection architecture, analysis sequencing, and stakeholder communication throughout.

4.2.Framework Application

Validity Gate. The Gate was applied to the full raw session dataset of 1,642 recorded sessions (including initial pilot tests which were excluded from wave1 analysis), implementing the Greenwald et al. (2003) D-score algorithm with extended engagement screening. Of these, 1,375 sessions reached full completion (Block 7); the remaining 267 partial sessions were excluded prior to gate evaluation. Of the 1,375 completers entering the gate, 12 sessions (0.9%) were excluded: 9 for fast responding (>10% of trials below 300ms) and 3 for high error rates (>30% of trials - Nosek et al., 2014), yielding a final valid dataset of 1,363 sessions across the full collection period to date. Following the Phase Comparability Check described below, the ~85 pilot sessions were excluded from substantive analysis. The Stage 1 analytical sample (N = 1,145) reflects the validated post-launch dataset at the point the Stage 1 analysis was conducted; the raw session figures reported here derive from a later data extract covering a longer collection window. The KMOP Stage 1 report (Pappas & Doufexi Kaplani, 2026) received the post-gate, post-phase-exclusion dataset as its analytical input, which is why that report records a 0% exclusion rate - no additional quality exclusions were required beyond those already applied by the Validity Gate at the infrastructure level. At 0.9%, the exclusion rate was substantially lower than the quality failure rates documented by Connors et al. (2020) in crowdsourced IAT samples, consistent with the higher engagement levels typically associated with organizationally-recruited and media-driven participant pools. This confirms that the Gate's

criteria were appropriately calibrated to the KMOP recruitment context, and that the Gate's output log - documenting every exclusion decision by criterion - provides the audit trail required for open science disclosure and funder review.

Phase Comparability Check. The Check revealed a statistically significant difference in mean D-score between pilot participants (N = 85, M = 0.08) and post-launch participants (N = 1,145, M = 0.17; t(94.7) = 2.10, p = .039). This is broadly consistent with evidence that different recruitment channels reach different populations (Nosek et al., 2002) - pilot participants in the KMOP study arrived through professional and researcher channels and showed markedly lower implicit bias scores than the broader post-launch pool.

The practical consequence of ignoring this difference is concrete: Had the pilot and post-launch datasets been merged without testing - the default practice in most web IAT deployments - the reported population mean would have been D = 0.15 rather than D = 0.17, a difference sufficient to shift the interpretive classification of the finding from within the slight bias range (D ≈ 0.15–0.35, following Project Implicit interpretive conventions; see Greenwald et al., 2003, p. 199, footnote 3, which maps IAT D-score ranges to Cohen's small/medium/large effect-size conventions) to the negligible-slight boundary. The Check caught this before it could affect the analysis. Accordingly, the pilot dataset was excluded from all Stage 1 analyses. This decision is documented, justified by the data, and eliminates a confound that most web IAT studies neither test for nor disclose.

Tiered Precision Architecture. The Architecture was applied across all demographic and psychographic subgroups of analytical interest: gender, generational cohort, educational attainment, employment sector, geographic residence, and psychographic profile. To ensure responsible reporting of subgroup findings, we established a five-band threshold framework applied to all demographic and psychographic subgroups prior to any analysis. Thresholds were determined a priori based on a combination of statistical power analysis, IAT-specific psychometric benchmarks, and data protection considerations.

The architecture classifies each planned subgroup analysis into one of five adequacy bands based on the per-group sample size available for the inferential claim being made.

Below N = 20 (Privacy), reporting of subgroup means is prohibited entirely, consistent with aggregate-data anonymisation principles in GDPR (Regulation (EU) 2016/679, Recital 162); the N < 20 threshold is a researcher-imposed safeguard rather than a formal regulatory requirement.

Between N = 20 and N = 79 (Unstable), subgroup estimates are considered too imprecise for substantive interpretation. Point estimates may be noted for internal monitoring but are excluded from published findings.

Tier 1 (Descriptive, N = 80-159) permits reporting of stable subgroup means with uncertainty quantification - confidence intervals, bootstrapped distributions - but precludes formal

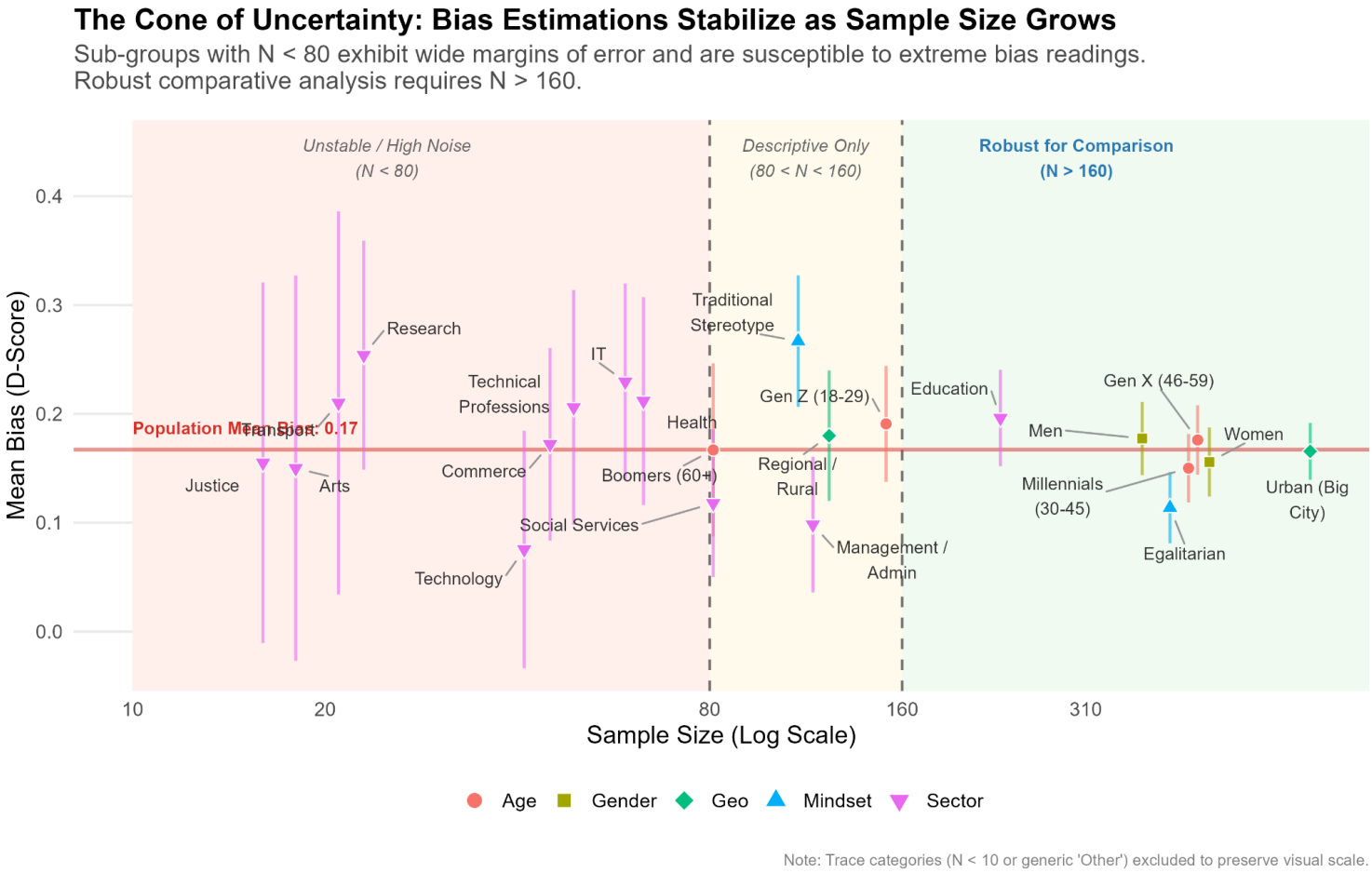
between-group comparative claims. The lower bound corresponds to the minimum N for 80% power in a one-sample t-test against zero (D = 0.15, SD ≈ 0.43, α = .05; Cohen, 1988), consistent with IAT effect-size benchmarks reported in Greenwald et al. (2003) and Nosek et al. (2002).

Tier 2 (Comparative, N = 160-309) permits formal two-group statistical comparisons, corresponding to approximately 80% power for detecting a small-to-medium between-group effect (Cohen's d ≈ 0.30) in an independent-samples t-test (Cohen, 1988).

Tier 3 (Benchmark, N ≥ 310) supports robust multi-predictor modeling and fine-grained subgroup decompositions. At this threshold, the 95% margin of error on a subgroup mean falls below ±0.05 D-score units - sufficient to distinguish meaningfully between adjacent bias magnitude categories (e.g., slight vs. moderate bias).

Tier thresholds are calibrated to the specific variance characteristics of the primary outcome measure and the effect sizes of theoretical and practical interest - a calibration that requires instrument-specific knowledge but follows a general logic applicable to any behavioral measure. All subgroup results reported in this paper are accompanied by their tier classification, and findings below Tier 1 are excluded from substantive interpretation.

At the time of Stage 1 analysis, the primary gender contrast (male/female) operated at Tier 2, supporting formal comparative claims. Several subgroups - particularly fine-grained age-by-sector and age-by-region interactions - operated at Tier 1, reported with descriptive estimates and confidence intervals but without formal comparison claims. The tier classification for each subgroup is disclosed explicitly in the Stage 1 report, making the precision status of every finding transparent to readers. This is a material improvement on standard practice in applied IAT reporting, where subgroup comparisons are frequently presented without reference to the adequacy of the underlying sample.



Note: Trace categories (N < 10 or generic 'Other') excluded to preserve visual scale.

Figure 2. Precision by Sample Size: Subgroup Bias Estimates Across the KMOP Gender-Career IAT (N = 1,145)

Each point represents a subgroup's mean implicit bias score (D-score) plotted against its sample size (log scale), with 95% confidence intervals. Background zones indicate analytical adequacy tiers: Unstable / High Noise (N < 80), Descriptive Only (80 < N < 160), and Robust for Comparison (N > 160). The red horizontal line marks the population mean bias (D = 0.17). Note: Subgroups with N < 10 or generic 'Other' categories excluded to preserve visual scale.

Temporal Stability Monitor. The Monitor tracked the daily mean D-score across the ten-week Stage 1 recruitment period. The resulting time series revealed the expected variation around the aggregate mean, with identifiable fluctuations corresponding to specific recruitment events - organizational newsletter distributions and targeted outreach campaigns producing short-lived compositional shifts that then resolved as the broader participant pool continued accumulating. The stability analysis confirmed that no single recruitment event produced a persistent shift in the aggregate estimate sufficient to require phase-segmented analysis. This finding is not self-evident and would not have been detectable without active monitoring. It supports the

interpretation of the Stage 1 dataset as a single analytical unit and strengthens the defensibility of the aggregate findings.

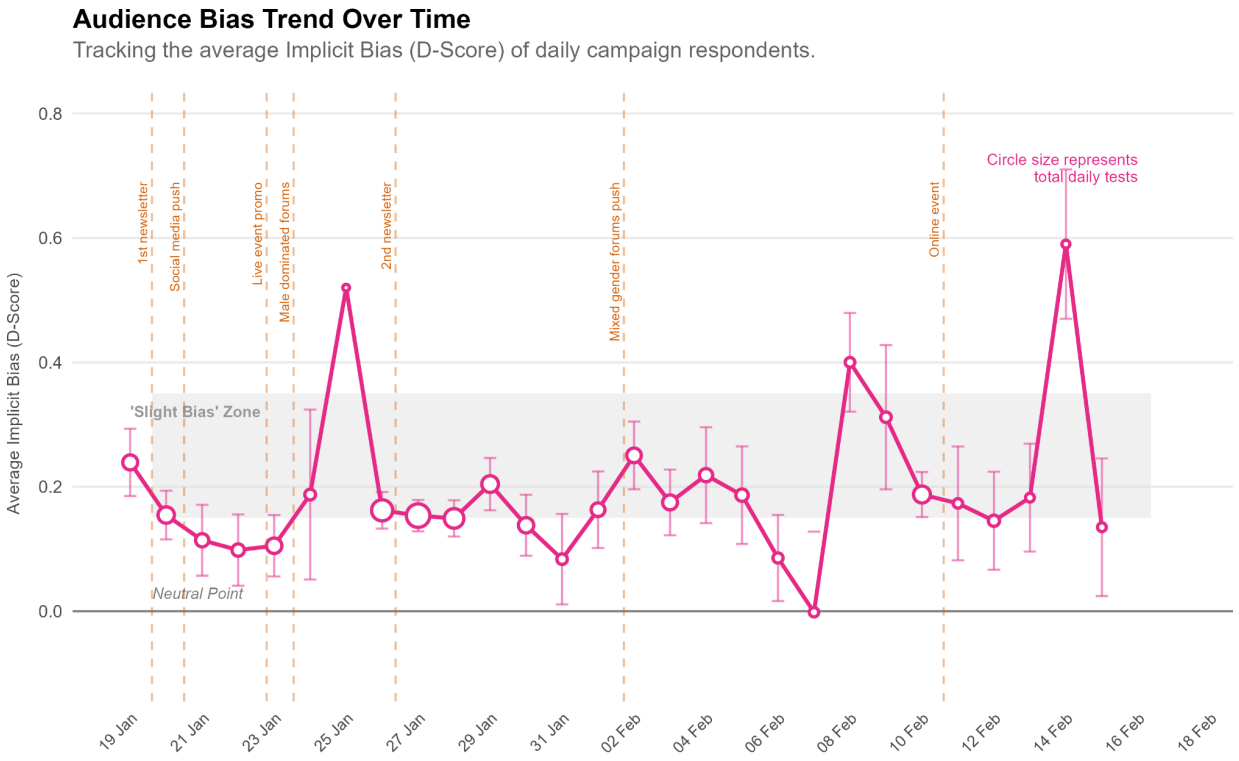


Figure 3. Temporal Stability of Daily Mean Implicit Bias Across the Recruitment Window (N = 1,145)

Daily mean D-score plotted across the full recruitment period (19 Jan – 17 Feb 2026), with standard error bars and bubble size proportional to daily test volume. The gray band indicates the 'Slight Bias' zone (D = 0.15–0.35 following Project Implicit interpretive conventions). Vertical dashed lines mark key recruitment campaign events (pseudonymised). Notable spikes following targeted outreach events — particularly channels with skewed audience demographics — illustrate the framework's capacity to detect sample composition drift in real time.

Recruitment Gap Analyzer. The Analyzer produced a prioritized subgroup gap map at the close of Stage 1 collection. This map identified the subgroups with the smallest gap to their next precision tier and translated those gaps into concrete recruitment targets by demographic profile. The map directly informed the design of the media outreach strategy for Stage 2 recruitment: communications were targeted to reach the participant profiles where precision gains were most achievable, rather than simply maximizing total N. The anticipated national media coverage of the Stage 1 findings is expected to drive Stage 2 participation above 2,000 total responses - a target calibrated by the Analyzer to the requirements of the preregistered confirmatory hypotheses.

```
======================================================
          GATE 7: RECRUITMENT WISHLIST
======================================================

>> QUICK WINS (Closest to the finish line)
# A tibble: 10 × 5
   Category   Group                  Current_N Needed Goal_Tier
   <chr>      <chr>                      <int>  <dbl>     <dbl>
 1 Generation Gen Z (18-29)                151      9       160
 2 Sector     Υγεία                         63     17        80
 3 Sector     Πληροφορική                   59     21        80
 4 Sector     Τεχνικά επαγγέλματα           49     31        80
 5 Sector     Εμπόριο                       45     35        80
 6 Sector     Τεχνολογία                    41     39        80
 7 Sector     Διοικητικές θέσεις           116     44       160
 8 Mindset    Traditional Stereotype       110     50       160
 9 Residence  Rural                        29     51        80
10 Sector     Έρευνα                        23     57        80

>> STRATEGIC TARGETS
# A tibble: 3 × 5
  Category   Group                  Current_N Needed Goal_Tier
  <chr>      <chr>                      <int>  <dbl>     <dbl>
1 Generation Gen Z (18-29)                151      9       160
2 Mindset    Traditional Stereotype       110     50       160
3 Generation Boomers (60+)                 81     79       160
```

Table 1. Recruitment Gap Analyzer Output: Subgroups by Distance to Next Adequacy Tier (N = 1,145)

Quick Wins: subgroups closest to their next adequacy tier threshold, ranked by participants needed. Strategic Targets: preregistered priority subgroups requiring targeted outreach. Goal Tier thresholds follow the Tiered Precision Architecture described in Section 3. Sector labels translated from Greek original: Υγεία = Health, Πληροφορική = IT, Τεχνικά επαγγέλματα = Technical Professions, Εμπόριο = Commerce, Τεχνολογία = Technology, Διοικητικές θέσεις = Management/Admin, Έρευνα = Research.

5. Implications for Applied IAT Research

5.1.Quality Assurance as Research Design

The standard conception of data quality assurance treats it as a post-collection activity: data are gathered, then cleaned, then analyzed. The Headway framework challenges this sequencing in a specific and consequential way. Each of the framework's five components is most valuable when integrated into the research design before collection begins - shaping instrument configuration, data architecture, and analytical planning in ways that determine what the eventual findings can credibly claim.

This integration changes what a research team can honestly communicate about its results. A study that has documented its exclusion criteria before collection begins (not after), confirmed the analytical comparability of its recruitment phases (not assumed it), constrained its subgroup inferences to statistically adequate cells (not overstated), monitored for sample drift (not ignored it), and anticipated its precision requirements in advance (not discovered gaps retrospectively) - that study produces findings that are qualitatively more defensible than one that did none of those things, even if the surface-level results look similar. The difference is not just academic. It determines whether findings survive peer scrutiny, whether they hold up in policy advocacy contexts, and whether they meet the open science disclosure standards required by Horizon Europe and analogous funders.

The chain of consequence is direct: Better-documented quality assurance produces more credible findings; more credible findings produce stronger grant applications, more defensible policy advocacy, and more durable institutional reputations. For organizations whose research outputs serve as evidence in funding competitions, public communications, or regulatory contexts, this is not a methodological nicety, but rather a strategic asset.

5.2.Research Partnerships and Horizon Relevance

The KMOP-Headway collaboration model is directly relevant to European research funding, particularly within Horizon Europe's emphasis on interdisciplinary partnerships, open science practices, and the active involvement of non-academic actors in research and innovation projects.

Horizon calls in thematic areas including gender equality (including Cluster 2: Culture, Creativity and Inclusive Society), health behavior, labor market inclusion, and social innovation increasingly require or prefer the use of behavioral measurement instruments - IATs, response-time tasks, digital trace measures - to quantify attitudes, biases, and behavior change. Delivering these instruments at scale, with documented quality assurance, requires exactly the combination of technical infrastructure and methodological expertise that the field literature identifies as absent from most non-academic deployments.

Headway addresses that gap. Non-academic partners with domain expertise and community reach - NGOs, think tanks, advocacy organizations, policy institutes - often lack the methodological infrastructure to meet the research quality standards required by academic-led consortia. By providing that infrastructure as a deployable service, Headway enables

substantive domain partners to participate in Horizon-quality research without building research engineering capacity in-house. The result is a partnership model where non-academic actors contribute what they are genuinely best positioned to contribute - community access, domain knowledge, communication reach, stakeholder relationships - while Headway provides the measurement and quality infrastructure those contributions require.

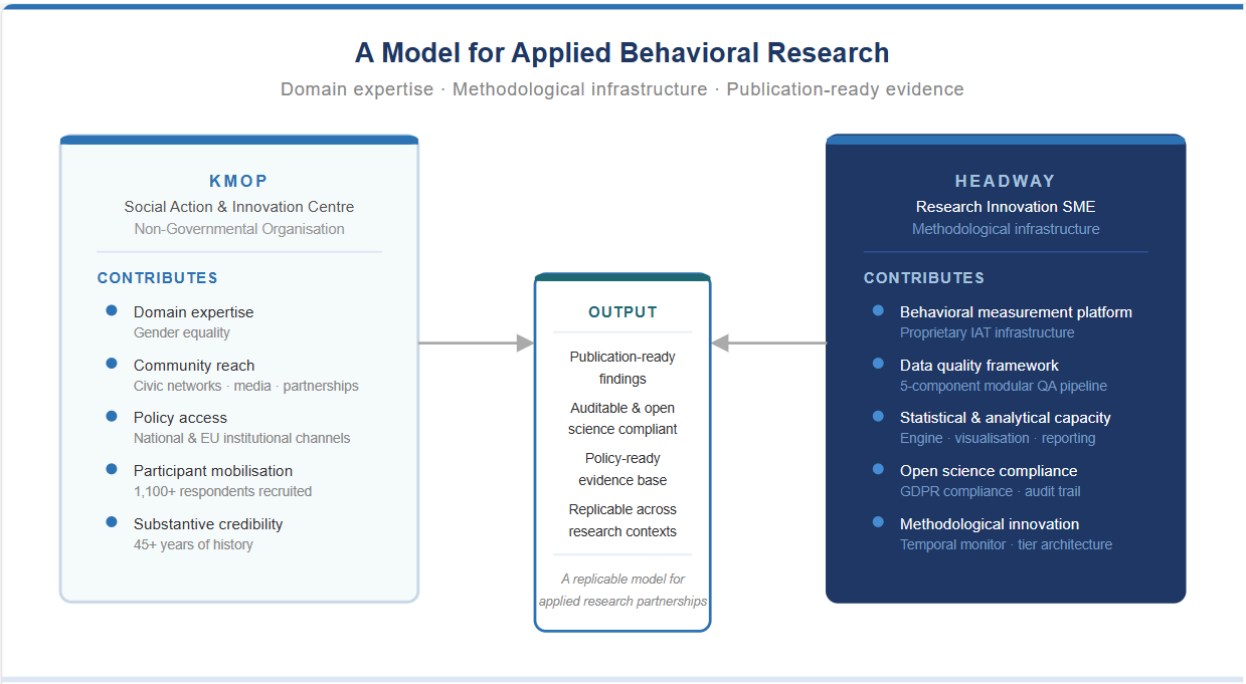


Figure 4. The KMOP–Headway Collaboration Model

A non-governmental organisation with domain expertise, community reach, and policy access (KMOP) partnering with a research innovation SME providing behavioral measurement infrastructure and analytical capacity (Headway). The collaboration demonstrates how complementary assets between applied research actors can produce publication-ready, open science-compliant evidence — and a replicable model for similar partnerships.

The Headway Data Quality Framework is available as a research infrastructure service for consortium partners requiring validated behavioral data collection and quality assurance. Teams considering the framework's application to their specific research context are encouraged to contact Headway directly.

5.3.Framework Generalizability and Future Directions

The five components described in this paper were developed in the context of IAT research, which serves as an illustrative implementation of the framework's broader logic. Their current implementation is optimized for the specific technical and statistical characteristics of the IAT - response latency distributions, D-score computation, trial-level exclusion logic - and the specific

challenges of open-recruitment web behavioral studies. However, the design principles underlying each component generalize across behavioral instruments: the validity gate logic applies to any response-time instrument; the phase comparability and temporal stability logic applies to any study with a multi-phase or extended collection window; the tiered precision architecture applies to any subgroup analysis plan; the recruitment gap analyzer applies to any study with defined precision requirements and a growing dataset. Tier thresholds and exclusion criteria are context-dependent and should be calibrated to the variance characteristics and effect sizes relevant to each instrument and research context.

Applicability Beyond the IAT

The following illustrates how each framework component maps onto research contexts beyond implicit association testing:

- **Survey-based attitude measures** (e.g. Likert-scale gender equality indices): Validity gating screens for straight-lining and inattentive responding; tiered precision restricts subgroup comparisons to adequately powered cells; temporal monitoring detects shifts in sample composition across extended fieldwork windows.
- **Experience sampling protocols** (e.g. repeated cross-sectional surveys tracking norm perceptions across policy intervention periods): Phase comparability confirms that early and late recruitment waves are compositionally equivalent; recruitment gap analysis identifies which participant profiles require targeted follow-up to support planned within-person analyses.
- **Response-time and cognitive load tasks** (e.g. response-time resume evaluation paradigms used in diversity and inclusion research): Validity gating implements RT distribution screening consistent with each task's psychometric requirements; temporal stability monitoring flags practice effects or platform changes that shift aggregate performance over time.

The field is moving toward standardized quality assurance and reporting frameworks for online behavioral research. Gagné and Franzen (2023) and Rodd (2024) have both emphasized the need for systematic documentation of data-quality procedures in online behavioral research. As such documentation standards continue to develop in the field, they will formalize expectations for what Headway's framework already implements operationally. Organizations that adopt the framework now are not just improving their current research - they are positioning themselves ahead of a reporting standard that Horizon and publishing outlets are likely to converge on.

Headway's roadmap includes the extension of this framework to survey-based behavioral measures, experience sampling protocols, and deliberative response instruments. As that extension proceeds, the framework will be formally validated across multiple instrument types and research contexts, providing a generalized quality assurance infrastructure for the European applied behavioral research landscape.

6. Conclusion

The applied behavioral research landscape is undergoing a structural shift. Organizations that once consumed research findings are now producing them - commissioning IATs, deploying behavioral instruments, and generating data that inform policy advocacy, public communication, and grant deliverables. The methodological infrastructure to support this shift has not kept pace.

This paper has presented the Headway Data Quality Framework: a modular, five-component system that operationalizes research-grade quality assurance for online behavioral research. Its components - Validity Gate, Phase Comparability Check, Tiered Precision Architecture, Temporal Stability Monitor, and Recruitment Gap Analyzer - address documented gaps in current web-based IAT practice, grounded in an extensive literature on web-based behavioral research quality. Their deployment in the KMOP Gender-Career IAT has demonstrated their practical value across a real large-scale study conducted under the conditions - open recruitment, heterogeneous devices, multi-phase collection - that the literature identifies as the most demanding.

The framework makes a concrete difference to what organizations can claim, document, and defend about their behavioral research findings. It meets the quality and transparency expectations of European research funders. And it enables substantive domain partners - NGOs, policy organizations, advocacy bodies - to participate in Horizon-quality research without bearing the full cost of building methodological infrastructure in-house.

The framework described in this paper is operational, documented, and available as a research infrastructure service. Organizations conducting online behavioral research in applied or consortium settings - and seeking quality assurance infrastructure that meets the transparency and rigor expectations of European funders - are invited to contact Headway to discuss its application to their specific research context.

7. References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N. Z., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bartmess, M. P., & Abreu, T. (2024). Maintaining Data Quality When Using Social Media for Recruitment: Risks, Rewards, and Steps Forward. Methodology, 20(4), Article e13839. https://doi.org/10.5964/meth.13839

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Carpenter, T. P., Pogacar, R., Pullig, C., Kouril, M., Aguilar, S., LaBouff, J., Isenberg, N., & Chakroff, A. (2019). Survey-software implicit association tests: A methodological and empirical analysis. *Behavior Research Methods*, 51(5), 2194–2208. https://doi.org/10.3758/s13428-019-01293-3

Chmielewski, M., & Kucker, S. C. (2019). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Connors, S., Spangenberg, K., Perkins, A. W., & Forehand, M. (2020). Crowdsourcing the Implicit Association Test: Limitations and best practices. *Journal of Advertising*, 49(3), 267–285. https://doi.org/10.1080/00913367.2020.1806155

Gagné, N., & Franzen, L. (2023). How to run behavioural experiments online: Best practice suggestions for cognitive psychology and neuroscience. *Swiss Psychology Open*, 3(1), 1. https://doi.org/10.5334/spo.34

Greenwald, A.G., Brendl, M., Cai, H. *et al.* Best research practices for using the Implicit Association Test. *Behavior Research Methods* 54, 1161–1180 (2022). https://doi.org/10.3758/s13428-021-01624-3

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Kochari, A. R. (2019). Conducting web-based experiments for numerical cognition research. *Journal of Cognition*, 2(1), 39. https://doi.org/10.5334/joc.85

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test

and intergroup behavior: A meta-analysis. American Psychologist, 74(5), 569–586. https://doi.org/10.1037/amp0000364

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101

Nosek BA, Bar-Anan Y, Sriram N, Axt J, Greenwald AG (2014) Understanding and Using the Brief Implicit Association Test: Recommended Scoring Procedures. PLoS ONE 9(12): e110938. https://doi.org/10.1371/journal.pone.0110938

Pappas, Y., & Doufexi Kaplani, M. E. (2026). *Gender, Career & Unconscious Bias in Greece: Findings from a Large-Scale Implicit Association Test Study (Wave 1)*. KMOP - Social Action and Innovation Centre. https://doi.org/10.5281/zenodo.18801509

Peer, E., Rothschild, D., Gordon, A. et al. Data quality of platforms and panels for online behavioral research. Behav Res 54, 1643–1662 (2022). https://doi.org/10.3758/s13428-021-01694-3

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript web experiments. *Behavior Research Methods*, *47*(2), 309–327. https://doi.org/10.3758/s13428-014-0471-1

Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should we stop looking for a better scoring algorithm for handling implicit association test data? Test of the role of errors, extreme latencies treatment, scoring formula, and practice trials on reliability and validity. *PLOS ONE*, 10(6), e0129601. https://doi.org/10.1371/journal.pone.0129601

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, 134, 104472. https://doi.org/10.1016/j.jml.2023.104472

Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2023). From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of Cognition*, 6(1), 13. https://doi.org/10.5334/joc.259