Published in final edited form as:
Anal Chem. 2021 October 19; 93(41): 13870–13879. doi:10.1021/acs.analchem.1c02621.

Nontargeted Analysis Study Reporting Tool: A Framework to Improve Research Transparency and Reproducibility

Katherine T. Peter,
U.S. National Institute of Standards and Technology, Charleston, South Carolina 29412, United States

Allison L. Phillips,
U.S. Environmental Protection Agency, Durham, North Carolina 27709, United States

Ann M. Knolhoff,
Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland 20740, United States

Piero R. Gardinali,
Institute of Environment and Department of Chemistry & Biochemistry, Florida International University, North Miami, Florida 33181, United States

Carlos A. Manzano,
Faculty of Science, University of Chile, 7750000 Nunoa RM, Chile; School of Public Health, San Diego State University, San Diego, California 92182, United States

Kelsey E. Miller,
U.S. Environmental Protection Agency, Durham, North Carolina 27709, United States

Manuel Pristner,

Corresponding Authors ktpeter@uw.edu, phillips.allison@epa.gov.

Author Contributions
These authors contributed equally. K.T.P. and J.R.S.—initial SRT development; K.T.P., A.L.P., A.M.K., and J.R.S.—study design; K.T.P., A.L.P., and J.R.S.—manuscript writing; K.T.P., A.L.P., A.M.K., and J.R.S.—SRT revisions; all authors—SRT reviews and manuscript editing.

ASSOCIATED CONTENT
Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c02621.
File 1: Reviewer demographics and expertise; summary of post-evaluation questionnaire responses; compiled external and self-reviews for each article; compiled V1, V2, and V3 scores across all articles; summarized relationship between reviewer assignments of numeric and color-based scores; summarized results of external reviewer evaluations with scores translated to final NTA SRT scoring system; and final SRT scoring system with score level descriptions and representative examples (PDF)
File 2: Original NTA SRT, detailed information about the eight articles reviewed, complete (blinded) article reviews, and follow-up questionnaire responses (XLSX)
File 3: Interactive spreadsheet version of the SRT (XLSX)
File 4: Interactive one-page PDF version of the SRT (PDF)

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c02621

Views expressed in this article are those of the authors and do not necessarily represent views or policies of the U.S. Environmental Protection Agency, U.S. NIST, U.S. Food and Drug Administration, or Agriculture and Agri-Food Canada.

Identification of commercial equipment, instruments, or materials in this research to adequately specify experimental procedures does not imply recommendation or endorsement by the aforementioned institutions, nor imply that the materials or equipment identified are necessarily the best available for the purpose.

The authors declare no competing financial interest.

Department of Food Chemistry and Toxicology, Faculty of Chemistry, University of Vienna, 1090 Vienna, Austria

Lyne Sabourin,
Agriculture and Agri-Food Canada, London, Ontario N5V 4T3, Canada

Mark W. Sumarah,
Agriculture and Agri-Food Canada, London, Ontario N5V 4T3, Canada

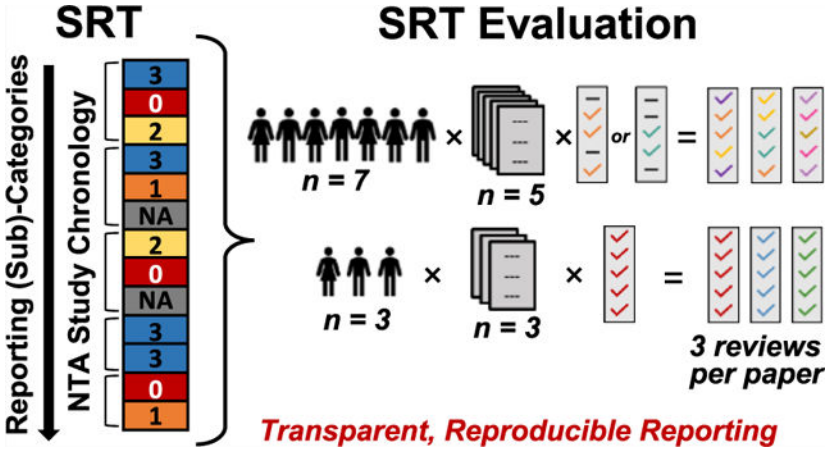
Benedikt Warth,
Department of Food Chemistry and Toxicology, Faculty of Chemistry, University of Vienna, 1090 Vienna, Austria

Jon R. Sobus
U.S. Environmental Protection Agency, Durham, North Carolina 27709, United States

Abstract

Non-targeted analysis (NTA) workflows using mass spectrometry are gaining popularity in many disciplines, but universally accepted reporting standards are nonexistent. Current guidance addresses limited elements of NTA reporting—most notably, identification confidence—and is insufficient to ensure scientific transparency and reproducibility given the complexity of these methods. This lack of reporting standards hinders researchers' development of thorough study protocols and reviewers' ability to efficiently assess grant and manuscript submissions. To overcome these challenges, we developed the NTA Study Reporting Tool (SRT), an easy-to-use, interdisciplinary framework for comprehensive NTA methods and results reporting. Eleven NTA practitioners reviewed eight published articles covering environmental, food, and health-based exposomic applications with the SRT. Overall, our analysis demonstrated that the SRT provides a valid structure to guide study design and manuscript writing, as well as to evaluate NTA reporting quality. Scores self-assigned by authors fell within the range of peer-reviewer scores, indicating that SRT use for self-evaluation will strengthen reporting practices. The results also highlighted NTA reporting areas that need immediate improvement, such as analytical sequence and quality assurance/quality control information. Although scores intentionally do not correspond to data/results quality, widespread implementation of the SRT could improve study design and standardize reporting practices, ultimately leading to broader use and acceptance of NTA data.

Graphical Abstract



Non-targeted analysis (NTA; also called “untargeted analysis” and “non-target screening”) experiments examine sample chemical composition beyond pre-defined targets to characterize unknown/understudied chemicals, classify samples, and discern trends otherwise missed by targeted analyses.[1] NTA studies employ various analytical hardware and software, with many using high-resolution mass spectrometry (HRMS). Diverse fields such as medicine, food science, and environmental health now leverage or retrospectively mine NTA data sets to provide innovative solutions to intractable problems.[2–6] Example applications include identifying early disease state biomarkers,[7–9] classifying samples for food safety evaluations,[10] and characterizing emerging contaminants in human-impacted environments.[11,12]

Varied research goals have driven rapid growth of complex NTA workflows that suit specific user needs. Community-wide guidance on study design, experimental analysis, and results communication is critical to the success of individual NTA applications.[13–15] However, most NTA research guidance is domain-specific,[16–19] addresses best practices for individual workflow components [e.g., identification confidence,[15,20,21] quality assurance/quality control (QA/QC)[22–24]], and is communicated by a stream of workshops and publications. This paradigm creates three fundamental challenges. First, guidance is siloed, only reaching target audiences within specific research domains—this limits cross-fertilization of research ideas and can yield duplication across NTA sectors. Second, NTA topical guidance can be complex, dynamic, and challenging to distill—this creates an entry barrier for new NTA researchers and can hamper high-impact study design, execution, and reporting. Third, a growing onus is placed on NTA practitioners to stay abreast of current recommendations—lack of adherence can cause research reporting/ reviewing inconsistencies, undermining the transparency and reproducibility of NTA publications.

Given these challenges, readily useable tools are needed that facilitate the dissemination of standardized guidance to all NTA practitioners. The ideal solution would: (1) succinctly cover critical aspects of NTA study design, execution, and reporting, with sufficient flexibility for use across NTA research domains; (2) be readily accessible to novice and experienced NTA researchers alike; (3) provide standardized metrics to support rapid, transparent, and consistent evaluation of merit in key study areas; (4) be amenable to

periodic updates, given research advancements and stakeholder input; and (5) facilitate subsequent community efforts to formally define NTA best practices, quality criteria, and performance standards.

Accordingly, we developed the NTA Study Reporting Tool (SRT)—a living framework for assessing the quality of NTA study reporting. The SRT evolved from efforts by the Benchmarking and Publications for Non-Targeted Analysis (BP4NTA; www.nontargetedanalysis.org)[25] working group to develop and disseminate information about NTA study design, results reporting, and quality assurance. While the BP4NTA reference content[26] includes suggestions to ensure NTA *research quality*, the SRT focuses on assessing NTA *reporting quality*. As such, the SRT was designed to aid NTA practitioners, reviewers, and editors evaluate the quality of research manuscripts and proposals from the perspective of comprehensive, reproducible, and transparent reporting. The BP4NTA reference content organization follows the SRT structure, complementing the stand-alone SRT (downloadable/fillable PDF/spreadsheet). Although we present a static version of the SRT herein, integration with the BP4NTA website allows continued evolution as the NTA research community's needs change. In this paper, we describe the development and intended applications of the SRT, present the results of an SRT evaluation (using recently published NTA studies to assess SRT efficacy), and highlight NTA reporting areas that need immediate improvement.

METHODS

To evaluate SRT validity for assessing NTA study reporting, 11 BP4NTA members evaluated 8 peer-reviewed articles.[27−34] These reviewers (authors of this article) were a geographically diverse (USA, Canada, and Europe) mix of government (n = 7) and academic (n = 4) researchers with a range of NTA experience (0.5−11 years), diverse research foci (environmental, food, and exposomics), and combined broad knowledge of chromatography and mass spectrometry instrumentation (Figure S1). The selected studies represented three general types—those focused on NTA performance evaluation,[30,31] development of data analysis methods for NTA ("NTA method development"),[28,32] and applications of NTA data ("NTA application")[27,29,33,34] —and a wide range of sample types, instrument/software platforms, and data analysis methods (Table S1). Most studies (six of eight)[28−33] were from the authors' publication records, facilitating permissions for inclusion and enabling comparison to "self-evaluations".

The evaluation effort yielded three complete evaluations of each paper (i.e., triplicate reviews per SRT sub-category per paper) (Figure 1). Seven "category reviewers" each evaluated five papers[30−34] across 1−3 SRT categories. Three additional "full SRT reviewers" (one of which represented a graduate student/advisor pair) each evaluated three papers[27−29] for all five SRT categories. This structure, partly borne of convenience, avoided overburdening reviewers while maximizing experience level and research expertise representation. It also enabled both sharply focused (category reviewers) and fit-for-purpose (full SRT reviewers) experiences with the SRT. SRT reviewers were blinded to specific reviewer identities when examining results but not to author identities while performing

reviews, because evaluated papers were previously published. Study authors performed a self-evaluation before seeing review results.

To determine the preferred scoring system, all reviewers used three versions: a qualitative 3-level system (Version 1 [V1]; *Yes/No/Not Applicable [NA]*); a color-coded 4-level system (Version 2 [V2]; *blue* = Yes, *yellow* = Needs more information, *red* = No, *gray* = *NA*); and a 7-level numeric system (Version 3 [V3]; *0*–5, *NA*; where 5 and *0* = all elements or no elements of relevant reporting were present, respectively) (Table S2). In all cases, *NA* indicated that the reviewer determined reporting on that topic to be outside the study scope. Although numerical scoring to communicate study quality is often discouraged,[35,36] the SRT focuses solely on reporting (not study outputs). Here, the numerical system facilitated statistical analyses and comparisons across reviewers and (sub-)categories. Reviewers were provided the scoring system descriptions given above (without further detail, to enable unbiased assessment) and instructed to only consider reporting quality (not study/data quality). Scores were assigned at the sub-category level to support specificity (and allow evaluation of score assignment at the category level). Reviewers provided open-ended rationales for scores and the relevant text/location in the publication (full evaluation results in Table S3). Triplicate non-*NA* reviewer scores were averaged to provide an overall reporting score for each paper in each sub-category. To gauge global reporting quality in each subcategory, median scores were calculated using article-specific average scores (n = 8). Reviewers also completed an open-ended, follow-up questionnaire about SRT content, coverage, usability, and scoring systems (Figure S2 and Table S4). Evaluation results and questionnaire responses were reviewed to identify trends or unexpected insights.

RESULTS AND DISCUSSION

NTA Study Reporting Tool.

The SRT (Table 1; original version used during evaluation in Table S2) is structured by sections, "categories", and *sub-categories*, with assigned scores and accompanying rationales based on reporting quality (not study quality) in each sub-category. The "Example Information to Report" column provides representative examples relevant to each sub-category, making the SRT accessible in a single-page, stand-alone format. Although reporting within articles may be dispersed throughout the text, SRT organization reflects overarching NTA study chronology. This structure enables SRT application not only during proposal/manuscript preparation and review but also during NTA study design by encouraging researchers to consider and incorporate important research elements.

In the Methods section, three categories ("Study Design", "Data Acquisition", and "Data Processing & Analysis") each contain three sub-categories. "Study Design" sub-categories (*Objectives & Scope, Sample Information & Preparation*, and *QC Spikes & Samples*) cover study aspects completed prior to sample analysis. "Data Acquisition" sub-categories (*Analytical Sequence, Chromatography*, and *Mass Spectrometry*) contain aspects related to instrumentation and sample analysis planning/execution. "Data Processing & Analysis" sub-categories cover the broad range of NTA data analysis efforts, including methods/approaches for *Data Processing* (e.g., initial data extraction/reduction), *Statistical & Chemometric*

Analysis (e.g., data evaluation, interpretation, and analyses other than identification efforts), and *Annotation & Identification*.

In the Results section, two categories ("Data Outputs", "QA/QC Metrics") each contain two sub-categories. "Data Outputs" sub-categories (*Statistical & Chemometric Outputs* and *Identification & Confidence Levels*) cover reporting of results that correspond to the methods detailed in *Statistical & Chemometric Analysis and Annotation & Identification* sub-categories. Intentional sub-category alignment across Methods and Results ensures harmonized study reporting. The "QA/ QC Metrics" sub-categories (*Data Acquisition* QA/QC and *Data Processing & Analysis* QA/QC) cover reporting on the quality, boundary, precision, and accuracy of data acquisition and data processing & analysis methods and results.[26]

SRT Scoring System Selection.

In considering review results (Figures 2, S3 and S4), we note that the SRT was not available when the evaluated studies were designed, executed, peer-reviewed, or published. Furthermore, scores do not correspond to scientific validity or data/results quality, as the SRT was designed to assess NTA *reporting* alone. Overall, reviewers used the entire range of available scores, and scores varied both across sub-categories (within a paper) and across papers (within a sub-category), potentially indicating the SRT supports evaluations that reflect current diversity in reporting practices.

Comparison of the three scoring systems indicated that V1 (qualitative) was the least information-rich, restricting visual interpretation of reporting quality (Figure 2). Reviewers least preferred V1 due to binary limitations of *yes*/*no* scoring and the challenge of determining an affirmative score threshold (Figure S2a). Thus, subsequent discussion focuses on V2 and V3. The V2 (color-coded) system offered a visual snapshot of reporting quality, though scoring variability across reviewers (for a given sub-category of a given paper) was not easily deciphered. The V3 (numerical) system allowed the most scoring nuance and conveyed information about reporting quality and reviewer variability.

Reviewers almost equally preferred the 4-level color-coded V2 and 7-level numeric V3 (Figure S2b). Importantly, reviewers noted that V2 lacked flexibility, whereas V3 offered so much flexibility that it could yield unnecessary scoring variance. Closer examination of V2 and V3 results indicated that the ends of each spectrum were aligned, with 5's typically paired with *blue* (n = 114/117, 97%) and *0*'s with *red* (n = 14/ 15, 93%) (Figure S5). Likewise, scores between *2* − *4* most commonly paired with *yellow* (2's: n = 16/19, 84%; 3's: n = 40/ 43, 93%; *4*'s: n = 66/89, 74%). However, *1*'s were almost evenly divided between *red* (n = 3/7, 43%) and *yellow* (n = 4/ 7, 57%). Provided rationales often explained subtle differences in reviewer interpretations of color/number assignments (e.g., some reviewers reserved *blue*/*5* for comprehensive reporting; others assigned *blue*/*4* if minor details were lacking). However, the overall SRT review meaning was conserved across V2 and V3 scores, with V3 allowing a greater differentiation of intermediate reporting quality.

Based on the equal preferences and relative consistency, we developed a hybrid 5-level color/number scoring system that balances the flexibility of V3 and the lower potential for

variability of V2. This final system uses *red = 0*, orange = *1*, *yellow = 2*, *blue = 3*, and *gray = NA*. To portray this hybridization, we mapped V3 scores onto the color scheme of the final scoring system (Figure 2) and translated V3 scores to the final 5-level (0 − 3, *NA*) scheme (Figure S6).

SRT Evaluation Results.

Overall Evaluation Results.—Comparing assigned scores by publication, most evaluated studies received both high and low sub-category scores (Figure 2, S3 and S4). None scored poorly in all sub-categories (average score < 3), indicating the SRT will advance NTA reporting without setting unrealistic expectations. The results suggest that certain NTA aspects have more rigorous and widely practiced reporting conventions (discussed further below). Two studies scored well in every SRT sub-category (Sobus et al. 2019 and Peter et al. 2018;[29,31] 11 and 10 of 13 average sub-category scores > 4, respectively), potentially reflecting "creator's bias" due to study author involvement in initial SRT development. Overall, evaluation results indicated the SRT supports reliable, objective appraisals of study reporting quality and could reasonably inform an editor in judging overall NTA article reporting quality.

SRT Consistency.—V2 and V3 sub-category scores were fairly consistent across reviewers, despite widely varied reviewer experience levels and expertise areas (Figure S1). Triplicate V2 scores were typically within the same or adjacent color block. With V3, 86% of triplicate scores occurred within a numerical range ≤ 2 (Figures 2, S3 and S4). The results suggest that the category-level (versus sub-category level) scoring would communicate limited specificity and nuance, as evidenced by considerable score disparities across certain sub-categories within a given category (Figure 3). Self-assigned scores (open squares, Figures 2, S3 and S4) generally fell within the range of peer-reviewed scores (72% within range; self-assessed *NAs* counted as within-range if ≥ 1 external reviewer assigned *NA*). This indicates the SRT facilitated a fair peer-assessment, and SRT use for self-evaluation will strengthen reporting practices.

In very limited instances, higher variability across reviewers was attributed to incorrect SRT application to examine perceived scientific rigor rather than reporting quality. For example, one reviewer gave Manzano et al. 2017[33] a *1/red* in *Identification & Confidence Levels*, with justification based on instrumentation and database type used for compound identification (Table S3d). In contrast, the other two reviewers assigned *4/yellow* and *5/blue*, listing rationales focused entirely on reporting quality. Because these biases occurred sporadically, no reviewer scores were excluded. Higher score variability across reviewers in certain sub-categories (Figure 3) highlighted areas of NTA reporting that may be inadequately defined and signaled a need for intentional enhancements to SRT content (described below) to improve scoring reliability. To enhance the understanding of SRT scoring metrics and enable reliable scoring in future applications, we developed definitions for each scoring level, with examples for the three more variable sub-categories (see SRT Usability; Tables 1, S5).

Study Design and Data Acquisition.—The SRT evaluation highlighted six sub-categories considered relevant for all reviewed publications (no *NA*s assigned; Figure 3). Relatively high scores (median score across papers ≥ 4) with limited variability across reviewers (score range ≤ 1 per sub-category for ≥ 5 of 8 studies) were observed for five of these sub-categories (Figure 4), including all three "Study Design" sub-categories and two "Data Acquisition" sub-categories (*Chromatography* and *Mass Spectrometry*). Two publications (McCord et al. 2017, Renaud et al. 2017)[28,34] received *QC Spikes & Samples* average scores < 2 because QC information was simply not reported. These study authors acknowledged this reporting gap during self-review, reflecting the importance of clear and transparent expectations for reporting practices. The overall excellent reporting and consistent reviewer evaluations in these areas (all within Methods) reflects the familiarity of fundamental methods reporting, which is deeply rooted in quantitative/targeted analysis studies.

In contrast, *Analytical Sequence* had the lowest median score within "Data Acquisition" (median = 3.7), with only 4 of 8 studies averaging a score ≥ 4 (Figure 4). Whether or not explicitly reported, analytical sequence is an established element of targeted/quantitative studies, where internal standards are used to correct for instrument drift, matrix suppression, and so forth. NTA is exceptionally sensitive to analytical sequence (e.g., sample run order, use of discrete batches) due to factors such as sample carryover and diminished/variable sensitivity.[24] Our results indicated that analytical sequence reporting has, to date, not been adequately emphasized by the NTA community. Although there are not yet standardized approaches to quantify and correct for variance, analytical sequence should be considered and reported in any NTA study, especially when performing statistical analyses on acquired peak areas. For example, one reviewer noted that analytical sequence reporting in Peter et al. 2018[29] "would have been useful to know given the comparative nature" of the study (Table S3h). In contrast, another reviewer noted that analytical sequence reporting in Tran et al. 2020[27] was less critical "because the authors are identifying compounds and not using statistical comparisons of their data" (Table S3g). These clear rationales helped clarify the importance of specific sub-category ratings. Notably, almost all authors (n = 7/8) self-scored their *Analytical Sequence* reporting as equal to or lower than the lowest external reviewer score (Figure 2), potentially indicating growing recognition of the importance of run order/ analytical batch information in NTA studies.

Data Processing and Identification.—Median scores for *Data Processing* and *Annotation & Identification* sub-categories were each 3.5, with average article scores ranging from 1.7−5. Likewise, *Identification & Confidence Levels* had a median score of 3.5, with average publication scores ranging from 2.5−5 (Figure 4). Arguably the crux and focus of many NTA studies, the large variability in reporting quality for these sub-categories reflects the challenge of sufficiently reporting the details of complex NTA workflows. Reviewers most often indicated incomplete reporting of software settings, selected thresholds, and identification methods (including incomplete library/ database descriptions). As one example of excellent reporting, Sobus et al. 2019[31] received average scores > 4.9 in both "Data Processing & Analysis" sub-categories (Figure 2) due to the inclusion of a supplementary table containing all software settings and detailed descriptions

of related procedures. However, their lower average *Identification & Confidence* Levels score (3.7) reflected a common reviewer critique regarding results reporting—missing MS/MS spectra. This may reflect challenges associated with MS/MS reporting in public spectral repositories, such as the need for data curation, interlaboratory comparability, as well as inconsistency across various instrumentation/settings, sample types, and matrices. Despite these challenges, reporting complete MS/MS spectral information is crucial for transparent results communication.

Results indicated that the SRT functioned well for evaluating reporting in the two studies that employed GC × GC separation coupled to low-resolution MS.[27,33] Based on this small-scale evaluation, the SRT appears robust across study types employing a wide range of hardware (chromatography and MS instruments) and software platforms, including those that rely on low-resolution databases (e.g., the NIST EI MS library) for compound annotation and identification. Overall, we anticipate that the SRT will encourage and remind NTA researchers to describe data processing and annotation settings in sufficient detail to ensure study reproducibility. Given the variety of vendor and open-source NTA software platforms/workflows, improved reporting (e.g., avoiding software-specific jargon and settings, including data extraction thresholds) will not only enhance reader comprehension of study methodology and aid less-experienced NTA researchers during method development, but also allow better translation across disparate workflows.

Statistical & Chemometric Analysis and Outputs.—The majority of *NA* assignments occurred in the *Statistical & Chemometric Analysis* and *Statistical & Chemometric Outputs* sub-categories (Figure 3). Reviewer rationales and subsequent discussions revealed that reviewers interpreted the terms “statistical analysis” and “statistical output” (used in the original SRT) narrowly, only assigning scores for reporting traditional statistical tests and chemometric approaches (e.g., differential analysis, hierarchical cluster analysis, and so forth) rather than broad data analyses (e.g., data summarization, evaluating variability, and so forth). This was particularly evident in the Renaud et al. 2017[28] evaluation, which investigated the selectivity (a statistical metric) of different LC-HRMS/MS modes for detecting pharmaceuticals in water but received two *NA*s in *Statistical & Chemometric Analysis* and three *NA*s in *Statistical & Chemometric Outputs*. Furthermore, despite intentionally pairing these sub-categories by name across Methods and Results sections, reviewers often assigned an *NA* to one statistics sub-category but a numerical score to the other (e.g., the Sobus et al. 2019[31] evaluation received two *NA*s in *Statistical & Chemometric Analysis* despite receiving three numerical scores in *Statistical & Chemometric Outputs*). These discrepancies likely indicated a need for clarification within the SRT, rather than true inconsistencies in statistical method and result reporting; accordingly, edits were made to the SRT (detailed below).

QA/QC Metrics.—The “QA/QC Metrics” sub-categories—*Data Acquisition QA/QC* and *Data Processing & Analysis QA/ QC*—had median scores of 3.3 and 3.5, respectively, with the most variable scores of any SRT sub-category (average perpaper scores ranged from 1 − 5). Studies that specifically included QA/QC samples and focused on NTA performance evaluation (e.g., Knolhoff et al. 2019, Sobus et al. 2019)[30,31] received 5’s from

all reviewers in both QA/QC sub-categories, indicating reviewers had a clear recognition of good reporting. These studies provide a model for thorough QA/QC reporting and demonstrate the study design planning needed to support robust QA/QC assessments. In contrast, median scores for the other six studies (i.e., those focused on NTA applications or NTA method development) were 3.2 and 2.7 for *Data Acquisition QA/QC* and *Data Processing & Analysis QA/QC*, respectively. *Data Processing & Analysis QA/QC* scores had a range ≥ 3 for 5 of 6 papers, indicating this aspect of reporting was particularly challenging to assess. However, reviewer rationales reflected a relatively clear understanding of which QA/QC aspects were missing from a given study (Table S3). Thus, the greater variability in scores was primarily attributed to inconsistent assessment of the importance of missing aspects, stemming from a lack of universal, clearly developed best practices for NTA QA/QC. In targeted/quantitative analytical studies, QA/QC metrics and figures of merit are key to method validation, acceptance, and defensibility. Accordingly, the development of robust best practices for conducting and reporting QA/QC in NTA studies will be paramount to advancing NTA applications into the regulatory arena. We are optimistic that use of the SRT will encourage study designs that incorporate QA/QC and improve its reporting in NTA studies, as well as help guide development of first-generation best practices and key performance metrics for NTA studies.

SRT Usability.

In the evaluation questionnaire, less than half of reviewers (n = 4/10) reported sole reliance on the “Example Information to Report” column; the remainder (n = 6/10, most with ≤ 5 years NTA experience) consulted both the Example column and the detailed BP4NTA reference content[26] (Figure S2c). Reviewers described consulting the reference content for specific (sub-) categories in which they had less knowledge depth, or as a starting point (rather than a continuous reference). This underscored the value of the SRT not only as a stand-alone tool for evaluating NTA study reporting for reviewers with established NTA knowledge bases, but also as a learning framework for less-experienced NTA researchers. Although time-demands likely depend on NTA experience level and review thoroughness, full SRT reviewers reported spending ≈1−2 h per article evaluation (including time reading the article). Reviewers noted an initial time cost associated with SRT use but indicated minimal time burden after familiarization. In fact, one reviewer reported that the SRT facilitated faster peer-review by quickly pinpointing present versus absent/lacking study details.

Considering reviewer questionnaire responses, topics raised during reviewer debriefing and observed variability in certain sub-categories, several “how-to” aspects of the SRT warranted greater clarity to improve its stand-alone functionality. New introductory text accompanying the SRT clarifies that the sections (Methods and Results) are intended solely to organize the SRT content and do not indicate the location in an article where information should be reported. Reporting may occur in the supporting information and/or via citations (although such citations should indicate which method details were applied in the current study to guide reader/reviewer attention). Additionally, (1) the Example column lists representative examples (not criteria or “must-haves”) and should not be used as a “checklist” in determining sub-category scores—indeed, a reviewer must determine whether

additional details not explicitly listed are also critical, and (2) the reviewer should consider both the overall study objectives and conceptual linkages across SRT sub-categories to determine whether a sub-category is applicable to a given paper. These two points are challenging, requiring reviewer expertise and discretion to select *NA* versus *0* if information is not reported (Tables 1, S5). The Rationale column provides space for communication between authors and reviewers about these potential gray areas, consistent with typical peer review processes.

Finally, we considered several scoring system usability aspects. Several reviewers assigned non-integer numbers for V3 scores. To avoid future confusion, the final version of the SRT uses drop-down options that limit reviewers to integer-only scores (*0* − *3* = red-orange/yellow-blue; *NA* = gray). Calculation of an overall reporting quality score and differential sub-category weighting were considered (and strongly debated). However, given score variability across sub-categories and the potential need for weighting according to different study objectives, we decided that these additions would obscure important nuance. Instead, we provided a plotting functionality in the spreadsheet version of the SRT (enabling quick visual comparison of multiple reviewer scores; Supporting Information File 3) and reserve judgement of overall assessment results for end-users. Finally, we provided descriptions of each scoring level in the final hybrid color-coded/numerical system, with representative examples for selected sub-categories (Tables 1, S5).

SRT Coverage.

Reviewers identified potential SRT cover-age gaps related to: sub-categories to reflect both basic statistics (for data evaluation/comparison) and chemometric analyses; the inclusion of manual compound annotation/identification efforts; and the use of low-resolution mass spectral databases (particularly for GC studies). Edits were made to address concerns with a final review by the entire study team and are reflected in the SRT in Table 1 (and in Supporting Information Files 3 and 4). We expect that the SRT will continue to evolve based on community feedback, and as improvements are made to data processing workflows, instrumentation, and software. A “live” SRT is hosted on the BP4NTA website, with a feedback portal (www.nontargetedanalysis.org/SRT). A BP4NTA sub-committee will regularly update the SRT to capture user feedback and major advancements in the field. Older SRT versions will remain downloadable.

CONCLUSIONS

Our assessment showed that the SRT offers a functional, valid framework to guide manuscript writing and evaluate reporting quality for key NTA study aspects. Reviewing NTA manuscripts and proposals is notoriously challenging, particularly given rapid growth in analytical capabilities, information-rich data sets, data analysis tools/approaches, and diverse NTA applications. The SRT can be implemented throughout the research process: (1) as a study design tool because it reminds researchers to consider and incorporate key research elements; (2) during manuscript or grant proposal submission as a self-appraisal of reporting quality; and (3) during publication and proposal peer-review to inform editors and decision makers. However, the SRT should not be the sole consideration in the

review process, as an excellent score does not reflect scientific merit (i.e., researchers could receive high marks for clear reporting on a fundamentally flawed analysis). Still, the SRT encourages scientific transparency—study designs are often impacted by practical considerations (e.g., limited field site access to replicate sample collections, chemical standard availability) that do not fundamentally detract from study utility but should be clearly acknowledged to allow accurate interpretation of data quality and study outcomes. As interest develops in NTA for regulatory and risk management applications, a parallel need grows for reliable and reproducible approaches to evaluate NTA study design and reporting; the NTA SRT addresses this need.

Our SRT evaluation considered a limited sample of NTA publications (though highly diverse in scope and utilized techniques). Given the small sample size, it is unclear whether observed trends apply to all relevant publications. Nevertheless, articles receiving high scores in specific sub-categories provide positive examples for quality reporting in each area. When administered on a community level, the SRT has the potential to uncover trends in study reporting habits of the NTA field at large. BP4NTA members have already and will continue to employ the SRT during preparation and peer review of NTA articles for scientific journals, improving the reporting quality of numerous articles.

The SRT is intentionally flexible, with accompanying infrastructure allowing adaptation to the evolving community needs. Creating a parallel tool for assessing NTA study/data quality that is universal to all NTA applications is currently not feasible. Development of innovative methods to analyze and apply NTA data will continue for the foreseeable future. We envision that a widespread adoption of the SRT will drive universal improvements to NTA reporting, thereby advancing the community's ability to determine the best practices and performance metrics that will ultimately underpin all NTA studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors thank Cuong Tran, James McCord, and co-authors for allowing evaluation of their articles. We thank Nathan Dodder and James McCord for self-evaluations of their work and helpful comments on the article. Paulina Piotrowski and Jonathan Mosley provided excellent critical article reviews. We thank BP4NTA members (www.nontargetedanalysis.org/membership-list), particularly Benjamin Place, Elin Ulrich, Seth Newton, Sara Nason, Jon Challis, Bowen Du, Andrew McEachran, Yong-Lai Feng, Natalia Quinete, Laszlo Tolgyesi, Ruth Marfil-Vega, and Miaomiao Wang. This work was performed while K.T.P. held a National Research Council Postdoctoral Fellowship at the U.S. National Institute of Standards and Technology (NIST).

■ REFERENCES

(1). Milman BL; Zhurkovich IK TrAC, Trends Anal. Chem 2017, 97, 179–187.

(2). Hollender J; van Bavel B; Dulio V; Farmen E; Furtmann K; Koschorreck J; Kunkel U; Krauss M; Munthe J; Schlabach M; Slobodnik J; Stroomberg G; Ternes T; Thomaidis NS; Togola A; Tornero V.Environ. Sci. Eur 2019, 31, 42.

(3). Sobus JR; Wambaugh JF; Isaacs KK; Williams AJ; McEachran AD; Richard AM; Grulke CM; Ulrich EM; Rager JE; Strynar MJ; Newton SR J. Exposure Sci. Environ. Epidemiol 2018, 28, 411–426.

(4). Crimmins BS; Holsen TM Non-targeted Screening in Environmental Monitoring Programs. In Advancements of Mass Spectrometry in Biomedical Research; Woods AG, Darie CC, Eds.; Springer International Publishing: Cham, 2019, pp 731–741. DOI: 10.1007/978-3-030-15950-4_43

(5). Gao B; Holroyd SE; Moore JC; Laurvick K; Gendel SM; Xie ZJ Agric. Food Chem 2019, 67, 8425–8430.

(6). Alonso A; Marsal S; Julià A.Front. Bioeng. Biotechnol 2015, 3, 23. [PubMed: 25798438]

(7). Pinto FG; Mahmud I; Harmon TA; Rubio VY; Garrett TJ J. Proteome Res 2020, 19, 2080–2091. [PubMed: 32216312]

(8). Favretto D; Cosmi E; Ragazzi E; Visentin S; Tucci M; Fais P; Cecchetto G; Zanardo V; Viel G; Ferrara SD Anal. Bioanal. Chem 2012, 402, 1109–1121. [PubMed: 22101423]

(9). Pleil JD; Wallace MAG; McCord J; Madden MC; Sobus J; Ferguson GJ Breath Res. 2019, 14, 016006.

(10). Knolhoff AM; Fisher CM Food Chem. 2021, 350, 128540.

(11). Newton SR; McMahen RL; Sobus JR; Mansouri K; Williams AJ; McEachran AD; Strynar MJ Environ. Pollut 2018, 234, 297–306. [PubMed: 29182974]

(12). Tian Z; Peter KT; Gipe AD; Zhao H; Hou F; Wark DA; Khangaonkar T; Kolodziej EP; James CA Environ. Sci. Technol 2020, 54, 889–901. [PubMed: 31887037]

(13). Hites RA; Jobst KJ Environ. Sci. Technol 2018, 52, 11975–11976. [PubMed: 30354076]

(14). Pourchet M; Debrauwer L; Klanova J; Price EJ; Covaci A; Caballero-Casero N; Oberacher H; Lamoree M; Damont A; Fenaille F; Vlaanderen J; Meijer J; Krauss M; Sarigiannis D; Barouki R; Le Bizec B; Antignac J-P Environ. Int 2020, 139, 105545.

(15). Rampler E; Abiead YE; Schoeny H; Rusz M; Hildebrand F; Fitz V; Koellensperger G.Anal. Chem 2021, 93, 519–545. [PubMed: 33249827]

(16). Sumner LW; Amberg A; Barrett D; Beale MH; Beger R; Daykin CA; Fan TW-M; Fiehn O; Goodacre R; Griffin JL; Hankemeier T; Hardy N; Harnly J; Higashi R; Kopka J; Lane AN; Lindon JC; Marriott P; Nicholls AW; Reily MD; Thaden JJ; Viant MR Metabolomics 2007, 3, 211–221. [PubMed: 24039616]

(17). Cajka T; Fiehn O.Anal. Chem 2016, 88, 524–545. [PubMed: 26637011]

(18). Brazma A; Hingamp P; Quackenbush J; Sherlock G; Spellman P; Stoeckert C; Aach J; Ansorge W; Ball CA; Causton HC; Gaasterland T; Glenisson P; Holstege FCP; Kim IF; Markowitz V; Matese JC; Parkinson H; Robinson A; Sarkans U; Schulze-Kremer S; Stewart J; Taylor R; Vilo J; Vingron M.Nat. Genet 2001, 29, 365–371. [PubMed: 11726920]

(19). Taylor CF; Paton NW; Lilley KS; Binz P-A; Julian RK Jr.; Jones AR; Zhu W; Apweiler R; Aebersold R; Deutsch EW; Dunn MJ; Heck AJR; Leitner A; Macht M; Mann M; Martens L; Neubert TA; Patterson SD; Ping P; Seymour SL; Souda P; Tsugita A; Vandekerckhove J; Vondriska TM; Whitelegge JP; Wilkins MR; Xenarios I; Yates JR 3rd; Hermjakob H.Nat. Biotechnol 2007, 25, 887–893. [PubMed: 17687369]

(20). Schymanski EL; Jeon J; Gulde R; Fenner K; Ruff M; Singer HP; Hollender J.Environ. Sci. Technol 2014, 48, 2097–2098. [PubMed: 24476540]

(21). Metabolomics Quality Assurance & Quality Control Consortium (mQACC). https://epi.grants.cancer.gov/Consortia/mQACC/ (accessed 12/22/2020).

(22). Schulze B; Jeon Y; Kaserzon S; Heffernan AL; Dewapriya P; O'Brien J; Gomez Ramos MJ; Ghorbani Gorji S; Mueller JF; Thomas KV; Samanipour S.TrAC, Trends Anal. Chem 2020, 133, 116063.

(23). Knolhoff AM; Premo JH; Fisher CM Anal. Chem 2021, 93, 1596–1603. [PubMed: 33274925]

(24). Broadhurst D; Goodacre R; Reinke SN; Kuligowski J; Wilson ID; Lewis MR; Dunn WB Metabolomics 2018, 14, 72. [PubMed: 29805336]

(25). Place BJ; Ulrich EM; Challis JK; Chao A; Du B; Favela KA; Feng Y-L; Fisher CM; Gardinali PR; Hood A; Knolhoff AM; McEachran A; Nason S; Newton S; Ng B; Nunez JR; Peter KT; Phillips AL; Quinete N; Renslow RS; Sobus J; Sussman EM; Warth B; Wickramasekara S; Williams AJ An Introduction to the Benchmarking and Publications for Non-targeted Analysis Working Group, 2021. Unpublished Work.

(26). Newton SR; Nason S; Williams AJ; Peter KT Benchmarking and Publications for Non-Targeted Analysis. www.nontargetedanalysis.org/ (accessed 12/29/2020).

(27). Tran CD; Dodder NG; Quintana PJE; Watanabe K; Kim JH; Hovell MF; Chambers CD; Hoh E.Chemosphere 2020, 238, 124677.

(28). Renaud JB; Sabourin L; Topp E; Sumarah MW Anal. Chem 2017, 89, 2747–2754. [PubMed: 28194977]

(29). Peter KT; Tian Z; Wu C; Lin P; White S; Du B; McIntyre JK; Scholz NL; Kolodziej EP Environ. Sci. Technol 2018, 52, 10317–10327. [PubMed: 30192129]

(30). Knolhoff AM; Kneapler CN; Croley TR Anal. Chim. Acta 2019, 1066, 93–101. [PubMed: 31027538]

(31). Sobus JR; Grossman JN; Chao A; Singh R; Williams AJ; Grulke CM; Richard AM; Newton SR; McEachran AD; Ulrich EM Anal. Bioanal. Chem 2019, 411, 835–851. [PubMed: 30612177]

(32). Warth B; Spangler S; Fang M; Johnson CH; Forsberg EM; Granados A; Martin RL; Domingo-Almenara X; Huan T; Rinehart D; Montenegro-Burke JR; Hilmers B; Aisporna A; Hoang LT; Uritboonthai W; Benton HP; Richardson SD; Williams AJ; Siuzdak G.Anal. Chem 2017, 89, 11505–11513. [PubMed: 28945073]

(33). Manzano CA; Marvin C; Muir D; Harner T; Martin J; Zhang Y.Environ. Sci. Technol 2017, 51, 5445–5453. [PubMed: 28453248]

(34). McCord J; Strynar M.Environ. Sci. Technol 2019, 53, 4717–4727. [PubMed: 30993978]

(35). Jüni P; Witschi A; Bloch R; Egger M.JAMA 1999, 282, 1054−1060. [PubMed: 10493204]

(36). Higgins J; Thomas J; Chandler J; Cumpston M; Li T; Page M; Welch V.Cochrane Handbook for Systematic Reviews of Interventions; Cochrane, 2021.

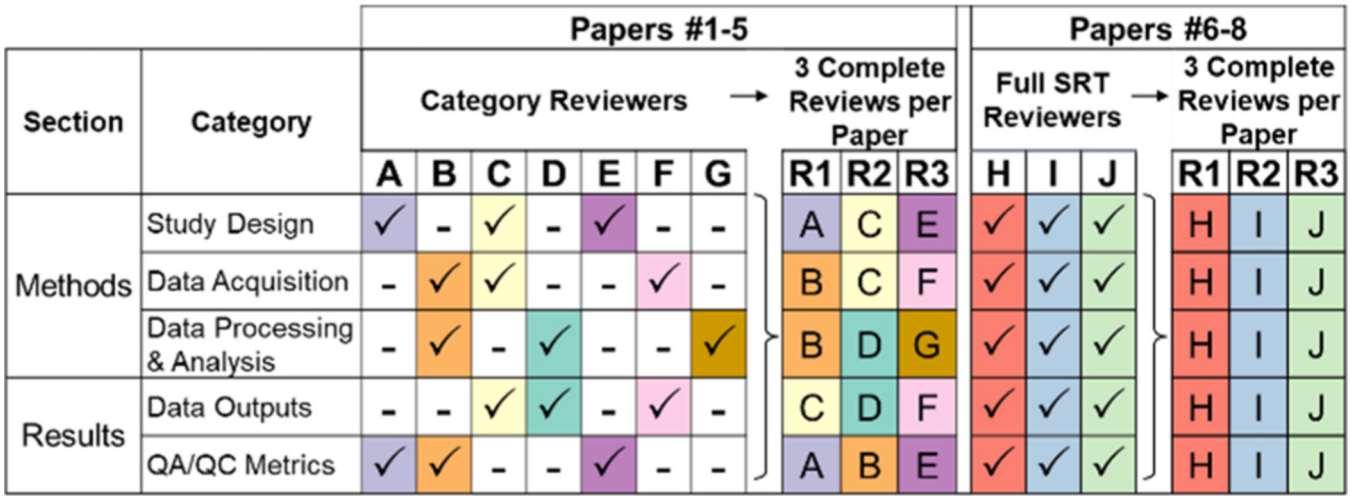


Figure 1.

Structure of the NTA SRT evaluation effort. Seven individuals (A−G, category reviewers) each reviewed five publications (Papers #1−5),[30−34] focusing on 1−3 SRT categories per article. Compiled category reviews yielded three complete reviews per paper. Three full paper reviewers (H−J; one of which represented a graduate student/advisor pair) each reviewed three publications (Papers #6−8),[27−29] focusing on all five SRT categories (yielding three complete reviews per paper). Reviewers did not review their own publications; authors were blinded to individual reviewer assignments. Publication authors ($n = 8$) self-reviewed their own papers after initial article reviews.

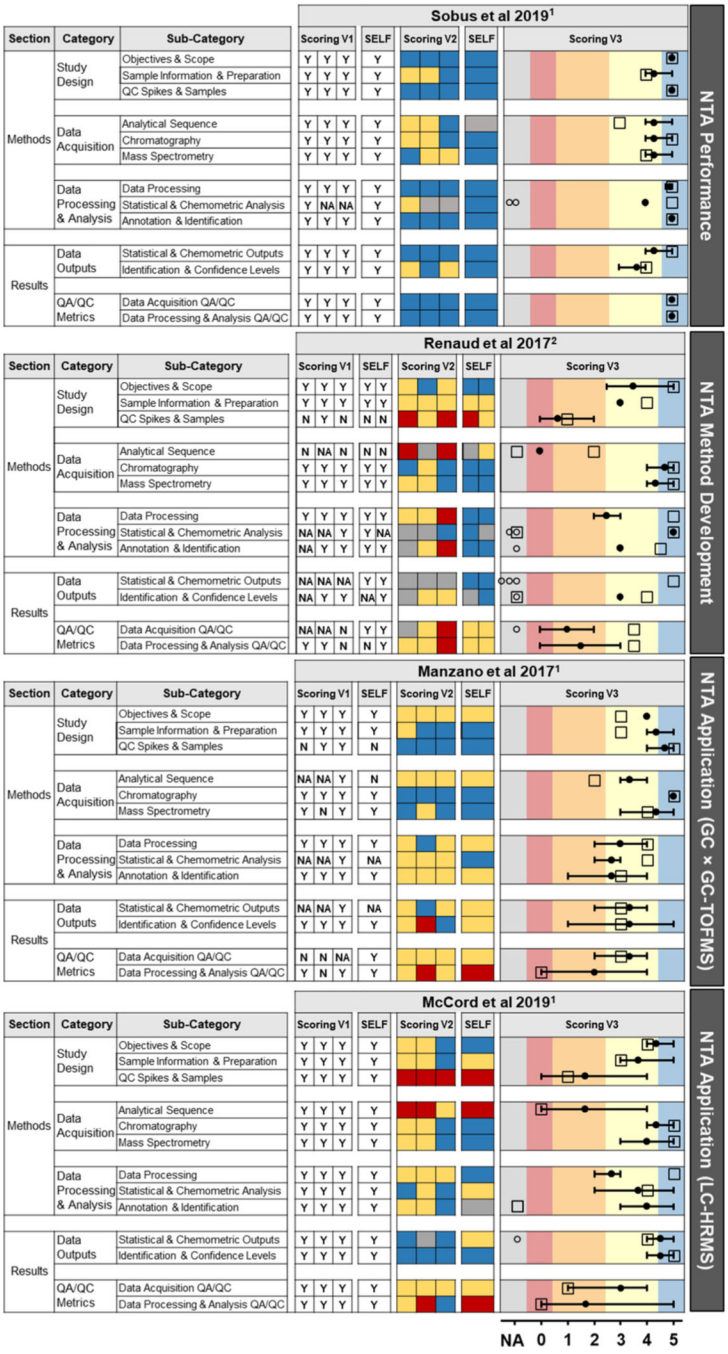


Figure 2.

External and self-review results for representative papers, including all three original scoring systems. In Scoring V3, filled circles show average external reviewer score, error bars show an external reviewer score range, open circles show external *NA* scores, and open squares show selfevaluation scores. Coloration behind Scoring V3 was applied after the evaluation (i.e., reviewers did not assign numeric scores with paired colors) by mapping the 6-level V3 scores (*NA*, 0 − 5) onto the final 5-level color scheme: *gray* = *NA*; *red* = 0 − < 0.5; *orange*

= 0.5 − < 2.5; *yellow* = 2.5 − < 4.5; *blue* = 4.5 − 5. Articles reviewed by [1]SRT category reviewers or [2]full SRT reviewers.

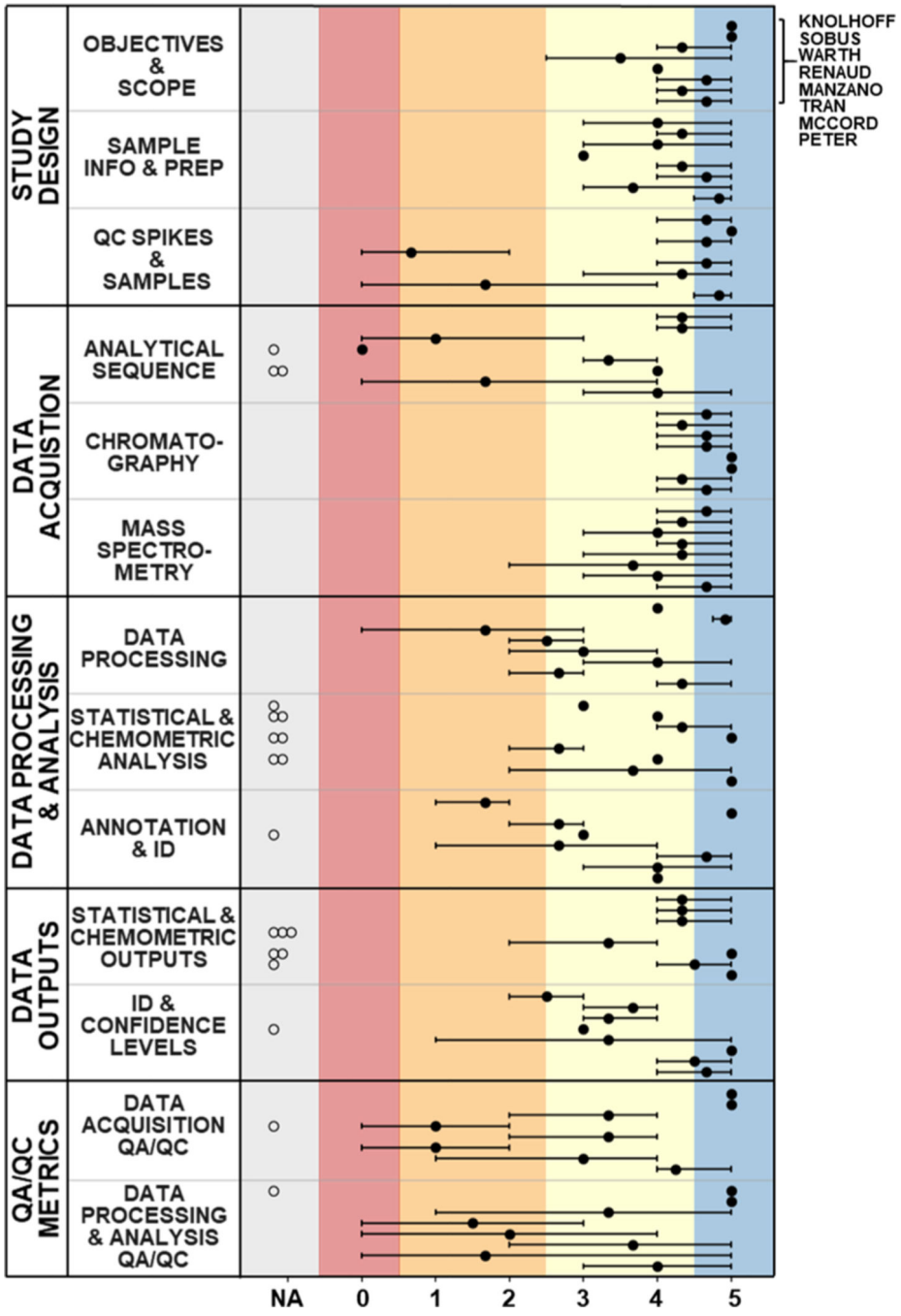


Figure 3.

V3 scoring results of external reviewer evaluations for all articles, grouped by SRT sub-category. All article scores are in the same order (from top to bottom, as noted at the figure top right) in each sub-category, and are grouped according to study type [NTA Performance—Knolhoff et al. and Sobus et al.;[30,31] NTA Method Development—Warth et al. and Renaud et al.;[28,32] NTA Application—Manzano et al. and Tran et al. (GC × GC-TOFMS),[27,33] McCord et al. and Peter et al. (LC-HRMS)].[29,34] Filled dots represent average reviewer scores, error bars represent reviewer score ranges, and open circles are

shown for *NA* scores. Coloration was applied after the evaluation process (i.e., reviewers did not assign numeric scores with paired colors) by mapping the 6-level V3 scores onto the final 5-level color scheme.

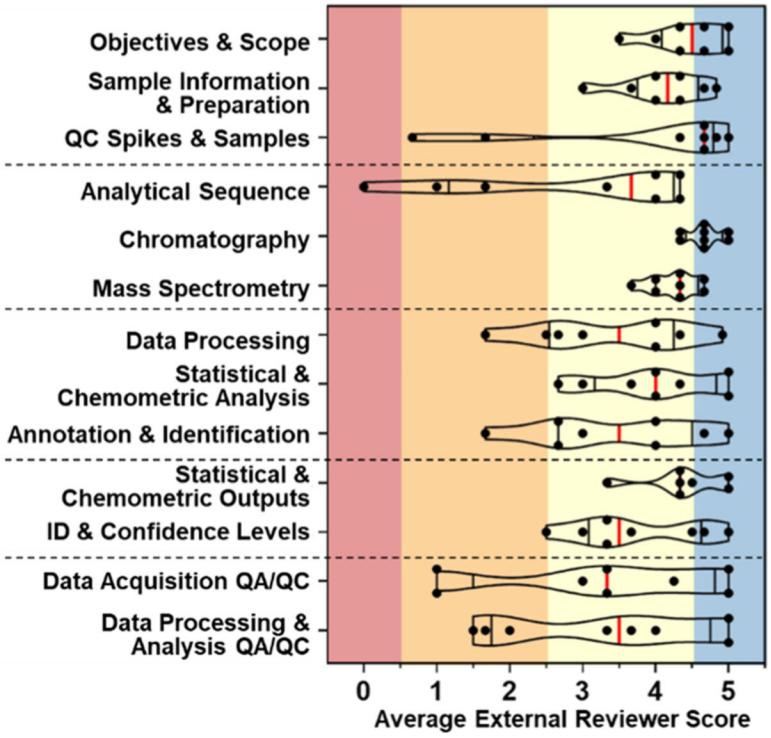


Figure 4.

Violin plot of average external reviewer scores (scoring system V3, *NA* scores excluded) for each article in each sub-category (dots). Median scores across all eight publications (vertical red lines) were calculated using article-specific average scores (vertical black lines show 25th and 75th percentile).

Table 1.

Current NTA Study Reporting Tool with Revisions Made to Reflect Feedback during the Evaluation Effort[a]

NTA Study Reporting Tool

Please read before using!

Purpose: This Tool was developed for use by NTA researchers and reviewers to assess the quality of NTA study reporting, and the resulting scores reflect solely whether the reporting is sufficiently complete and transparent (based on current, best available understanding of the environmental, food, and exposomics NTA communities). The Tool is not intended for evaluation of the quality of the study or resulting data.
We also encourage two supplementary uses of the Tool: 1) to guide study design - by considering what should be reported, a researcher is inherently encouraged to incorporate the necessary aspects into their study design, and 2) as a starting point for (or portal to) relevant reference content and resources, which are available via the BP4NTA website (www.nontargetedanalysis.org).

Notes & Guidance: The "Example Information to Report" column provides a brief list of representative items relevant to each sub-category - not all are required or necessary for every study. Researchers and reviewers should use their expertise and discretion to determine which points pertain to a given study, and whether additional details not explicitly listed are also critical to report. Additionally, certain sub-categories may not be relevant to a given study (hence the option to select "NA") or may be less critical to the overall quality and completeness of reporting. To evaluate these aspects, we strongly encourage users to both consider the study type and objectives (e.g., method development, performance evaluation, field application), as well as conceptual linkages across subcategories (e.g., between Statistical Analysis and Statistical Outputs). Please also note that the Sections (Methods and Results) are not intended to indicate the location in a manuscript where the information is reported - a user should consider the manuscript in its entirety (including any supporting documents and/or citations). We also encourage reviewers to include a rationale, so that authors/researchers may readily address concerns.

Scoring: *NA* = not applicable (gray); *3* (blue) is the highest score and *0* (red) is the lowest.

NA	Not relevant to study	0	No relevant elements reported	1	Some elements reported Major improvement needed	2	Most elements reported Minor improvement needed	3	All relevant elements reported

Section	Category	Sub-Category	Example Information to Report	Score	Rationale
Methods	Study Design	Objectives & Scope	• Study goals and hypotheses • Scope of the study with respect to use of NTA / suspect screening • Expected chemical coverage of approach and potential limitations	Scores selected from drop-down menu for each sub-category NA 0 1 2 3	Space for reviewer to explain assigned score in each sub-category
		Sample Information & Preparation	• Sample collection/replication, handling/storage, preparation, extraction, & clean-up methods (and related QA practices) • Intended use of samples (e.g., method development, compound identification, etc.) • Development and intended use of blanks		
		QC Spikes & Samples	• Development of QC spikes/samples (e.g., isotopically labeled standards/spikes, native standard spikes, matrix pools) • Intended use of QC spikes/samples (e.g., to monitor instrument performance, data normalization, etc.)		
	Data Acquisition	Analytical Sequence	• Sample randomization and use of replicate injections • Inclusion of blanks and QC samples in the acquisition sequence • Information about single vs. multiple analytical batches		
		Chromatography	• Instrument specifications • Method settings (e.g., column/guard, mobile phases, gradient, injection techniques)		
		Mass Spectrometry	• Instrument specifications • Instrument calibration and/or tuning procedures • Method settings (e.g., acquisition parameters, such as polarity, resolution, data-dependent vs. data-independent)		
	Data Processing & Analysis	Data Processing	• File conversion information (e.g., to open-source format, centroiding) • Software program(s) used • Workflow steps (e.g., peak picking, RT calibration, alignment, gap filling) and settings • Feature detection thresholds (e.g., replicate detection criteria; min height, area, or S/N levels; comparison to occurrence/abundance in blanks) • Data correction or normalization methods (e.g., peak area/height normalization or scaling, blank subtraction)		
		Statistical & Chemometric Analysis	• Software programs(s)/package(s) used & samples/sample groups to which analyses were applied • Basic statistical analysis method goals (e.g., summarize data, evaluate variability, hypothesis testing), type (e.g., Wilcoxon rank sum test, Chi-square test), assumptions, and settings/thresholds • Chemometric analysis method goals (e.g., prioritize features, compare/classify samples, evaluate relationships between features), type (e.g., differential analysis, hierarchical clustering, dimensionality reduction), assumptions, and settings/thresholds		
		Annotation & Identification	• Software program(s) used (or description of manual annotation/identification efforts) • Libraries and databases used (including details such as chemical coverage, resolution, metadata inclusion; information about in-house databases) • Workflow steps (e.g., formula assignment, suspect screening, MS/MS spectral interpretation or library matching) • Workflow methods & settings (e.g., formula prediction method, scoring algorithms; mass error/RT tolerances, accepted match scores)		
Results	Data Outputs	Statistical & Chemometric Outputs	• Basic statistical outputs (e.g., adj. p-values, standard deviations, test statistics) • Results of chemometric analyses (e.g., reported classifications/groupings of features or samples, observed trends in the data) • Visuals/plots (e.g., Venn diagrams, heatmaps, clustering dendrograms, volcano plots, network diagrams, PCA and loading plots) • New statistical metrics, algorithms, packages, and/or scripts		
		Identification & Confidence Levels	• Reported identifications and associated confidence levels (e.g., levels described by Schymanski et al., *ES&T*, 2014) • Supporting data for annotation/identification (e.g., formula match scores, fine isotope pattern, retention time match, MS/MS match scores, source of MS/MS spectra) • For features with lower confidence IDs, (i.e., not standard-confirmed), proposed tentative structures and other annotated data • Semi-quantification or quantification data • Exported MS/MS spectra (e.g., as a library, database, or deposition into online repository)		
	QA/QC Metrics	Data Acquisition QA/QC	• Quality: Adherence to QA/QC protocols for sample preparation and data acquisition • Boundary: Description of the potential impacts of methods (sample prep, chromatographic, MS) on observable chemical space • Accuracy: Reported chromatographic and mass accuracy • Precision: Variability of observed retention time, precursor mass error, and abundance		
		Data Processing & Analysis QA/QC	• Quality: Outcomes of QC checks along the data processing & analysis workflow • Boundary: Impact of data processing & analysis method(s) on observed chemical space, observed limits of detection/ID • Accuracy: Performance measures (True Positive Rate, False Positive Rate, etc.) for known compounds or samples with known classification • Precision: Reproducibility/repeatability of performance measures for known compounds or samples with known classification; Calculations such as False Discovery Rate, F1 score, etc.		

[a]The SRT is available for download as both Excel and PDF files on the BP4NTA website (www.nontargetedanalysis.org/SRT), with active scoring drop-down menus. These online versions (provided here as Supporting Information Files 3 and 4) will be updated regularly, based on community feedback.