Taylor & Francis
Taylor & Francis Group

Evidence-Based Toxicology

ISSN: 2833-373X (Online) Journal homepage: www.tandfonline.com/journals/tebt20

Protocol: Testing the Performance of INVITES-IN, A Tool for Assessing the Internal Validity of *In Vitro* Studies

Gro H. Mathisen, Gunn E. Vist, Paul Whaley, Richard A. White, Trine Husøy, Heather M. Ames, Anna Beronius, Emma Di Consiglio, Ingrid Druwe, Thomas Hartung, Sebastian Hoffmann, Carlijn R. Hooijmans, Kyriaki Machera, Pilar Prieto, Joshua F. Robinson, Erwin Roggen, Andrew A. Rooney, Nicolas Roth, Eliana Spilioti, Anastasia Spyropoulou, Olga Tcheremenskaia, Emanuela Testai, Mathieu Vinken & Camilla Svendsen

To cite this article: Gro H. Mathisen, Gunn E. Vist, Paul Whaley, Richard A. White, Trine Husøy, Heather M. Ames, Anna Beronius, Emma Di Consiglio, Ingrid Druwe, Thomas Hartung, Sebastian Hoffmann, Carlijn R. Hooijmans, Kyriaki Machera, Pilar Prieto, Joshua F. Robinson, Erwin Roggen, Andrew A. Rooney, Nicolas Roth, Eliana Spilioti, Anastasia Spyropoulou, Olga Tcheremenskaia, Emanuela Testai, Mathieu Vinken & Camilla Svendsen (2023) Protocol: Testing the Performance of INVITES-IN, A Tool for Assessing the Internal Validity of *In Vitro* Studies, Evidence-Based Toxicology, 1:1, 2293289, DOI: 10.1080/2833373X.2023.2293289

To link to this article: https://doi.org/10.1080/2833373X.2023.2293289

EVIDENCE-BASED TOXICOLOGY
2023, VOL. 1, NO. 1, 2293289
https://doi.org/10.1080/2833373X.2023.2293289

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Protocol: Testing the Performance of INVITES-IN, A Tool for Assessing the Internal Validity of *In Vitro* Studies

Gro H. Mathisen[a], Gunn E. Vist[a,b], Paul Whaley[a,c], Richard A. White[a,d], Trine Husøy[a,e], Heather M. Ames[a,b], Anna Beronius[f], Emma Di Consiglio[g], Ingrid Druwe[h], Thomas Hartung[i,j], Sebastian Hoffmann[k,l], Carlijn R. Hooijmans[m], Kyriaki Machera[n], Pilar Prieto[o], Joshua F. Robinson[p], Erwin Roggen[q], Andrew A. Rooney[r], Nicolas Roth[s,t], Eliana Spilioti[n], Anastasia Spyropoulou[n], Olga Tcheremenskaia[g], Emanuela Testai[g], Mathieu Vinken[u] and Camilla Svendsen[a,v]

[a]Norwegian Scientific Committee for Food and Environment, Norwegian Institute of Public Health, Oslo, Norway; [b]Division for Health Services, Norwegian Institute of Public Health, Oslo, Norway; [c]Lancaster Environment Centre, Lancaster University, Lancaster, UK; [d]Department of Method Development and Analytics, Norwegian Institute of Public Health, Oslo, Norway; [e]Department of Food Safety, Norwegian Institute of Public Health, Oslo, Norway; [f]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; [g]Environment & Health Department, Italian National Institute of Health (ISS), Rome, Italy; [h]Office of Research and Development, Center for Public Health and Environmental Assessments, United States Environmental Protection Agency, Research Triangle Park, NC, USA; [i]Center for Alternatives to Animal Testing (CAAT), Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD, USA; [j]CAAT Europe, University of Konstanz, Konstanz, Germany; [k]Evidence-based toxicology collaboration (EBTC), Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD, USA; [l]seh Consulting + Services, Paderborn, Germany; [m]Department of Anesthesiology, Pain and Palliative Care, Radboud University Medical Centre, Nijmegen, Netherlands; [n]Laboratory of Toxicological Control of Pesticides, Scientific Directorate of Pesticides' Control and Phytopharmacy, Benaki Phytopathological Institute, Kifissia, Greece; [o]European Commission, Joint Research Centre (JRC), Ispra, Italy; [p]Department of Obstetrics, Gynecology & Reproductive Sciences, University of California, San Francisco (UCSF), USA; [q]3Rs Management and Consulting ApS, Lyngby, Denmark; [r]Division of Translational Toxicology, National Institute of Environmental Health Sciences, DurhamNC, USA; [s]Department of Pharmaceutical Sciences, University of Basel, Basel, Switzerland; [t]Swiss Centre for Applied Human Toxicology (SCAHT), Basel, Switzerland; [u]Department of Pharmaceutical and Pharmacological Sciences, Vrije Universiteit Brussel, Belgium; [v]Department of Chemical Toxicology, Norwegian Institute of Public Health, Oslo, Norway

ABSTRACT
A tool for evaluation of internal validity of *in vitro* studies called INVITES-IN is currently under development. The tool is designed specifically for cell culture studies.
This protocol describes the testing of the performance of INVITES-IN. By performance, we mean the extent to which results of using INVITES-IN are the same for different users (consistency), the amount of time and cognitive effort it takes to apply INVITES-IN (assessor workload), the precision and potential for systematic error in results of applying INVITES-IN (accuracy), and how easy it is to use INVITES-IN (user experience).
The participants in the user testing will be representative for the expected end-users of INVITES-IN which are persons preparing literature reviews including *in vitro* studies (e.g. in the context of chemical hazard and risk assessments or drug development). All end-users are expected to have experience with *in vitro* methods.
Data collected from the performance testing will be used for further refinement and development of the release version of INVITES-IN.

Abbreviations: ICC: Intraclass Correlation Coefficient; MOE: margin of error; NAM: new approach methodologies; PG: project group; RoB: risk of bias; SAG: scientific advisory group; SQ: signalling question

ARTICLE HISTORY
Received 5 September 2023
Revised 14 November 2023
Accepted 23 November 2023

KEYWORDS
Cell culture; evidence-based toxicology; risk of bias; user testing

CONTACT Gro H. Mathisen gro.haarklou.mathisen@vkm.no

Supplemental data for this article can be accessed online at https://doi.org/10.1080/2833373X.2023.2293289.
This manuscript was accepted for publication by the handling editor Olwenn Martin after 1 round of editorial evaluation and 1 round of peer-review evaluation. The evaluation reports for the manuscript can be found at https://doi.org/10.5281/zenodo.10037725. Preprint versions of the manuscript and author responses to comments in the evaluation reports can be found at https://doi.org/10.5281/zenodo.10119424.

1. Introduction

We are developing a tool for assessing the internal validity of *in vitro* studies called INVITES-IN (detailed description of the creation of this tool is available in the protocol by Svendsen et al. (2023)). INVITES-IN will be designed specifically to be applied to eukaryotic cell culture studies including cell lines, primary cells, co-cultures, monolayer, and 3-D cell systems. As we are not able to directly measure internal validity, INVITES-IN will be designed to address the risk for introduction of different types of bias (risk of bias [RoB]) to the cell culture studies. In this protocol, we describe how we will test the performance of INVITES-IN. The participants in the user testing will be representative for the expected end-users of INVITES-IN which are persons preparing literature reviews including *in vitro* studies e.g. in the context of chemical hazard and risk assessments or drug development. This includes persons having practical experience from carrying out *in vitro* studies, and/or experience with evaluating evidence from *in vitro* studies.

INVITES-IN will consist of two elements: a set of signalling questions addressing different sources of bias (e.g. selection bias, performance bias, and detection bias) that may be of importance for the internal validity of the study, and a guidance document containing step-by-step instructions explaining how to assess the degree of bias through the rating of the signalling questions. The outcome of the three studies described in the protocol for the creation of INVITES-IN (Svendsen et al. 2023) will determine which bias domains and items that will be included in INVITES-IN. The users of INVITES-IN will answer the signalling questions in order to determine whether bias might have been introduced. For each signalling question, the criteria that will be used for assessing the risk for introduction of bias will be described in the guidance document. A signalling question is judged as being low risk for introduction of bias when all criteria for that question are fulfilled.

The creation of INVITES-IN follows the framework of general principles for developing quality assessment tools as suggested by Whiting et al. (2017). Four studies will be performed to develop, test, refine and issue the release version of the tool. The first three studies are described by Svendsen et al. (2023). In the first study, the relevance of bias domains and items for *in vitro* studies were interpreted through focus group discussions. In the second study, the importance and impact of the relevant bias items for *in vitro* studies will be evaluated and ranked using a modified Delphi method. In the third study, the beta version of INVITES-IN will be created, based on the outcome of the first two studies. With the fourth study, the performance testing, we aim to identify how well the tool works in practice to identify needs for improvement that should be considered when creating the release version of INVITES-IN. The current protocol describes the fourth study.

1.1 Objective

The objective of this study is to test the performance of INVITES-IN as a tool for assessing internal validity of cell culture studies to collect information for further refinement and development of the release version of INVITES-IN. The following will be evaluated:

1. Consistency, in terms of the extent to which results of using INVITES-IN are the same for different users.
2. Assessor workload, in terms of the amount of time and cognitive effort it takes to apply INVITES-IN.
3. Accuracy in the application of INVITES-IN to evaluate cell culture studies, in terms of the precision against the gold-standard application and potential for systematic error in results.
4. User experience in applying the tool, in terms of how easy it is to use INVITES-IN.

Testing the generalisability of INVITES-IN, i.e., how well the tool performs for assessment of the internal validity of other studies than cell culture studies, does not fall under the scope of this study.

1.2 Project governance

The development of INVITES-IN is part of the project “Next Generation Risk Assessment in Practice” (VKM 2023). This project is led by the Norwegian Institute of Public Health represented by the Norwegian Scientific Committee for Food and Environment, and the partners involved in this project are the Benaki Phytopathological Institute, the Italian Institute of Health, and the University of Basel. The project is part of the European Partnership for the Assessment of Risks from Chemicals (PARC) [Project 101057014]. PARC aims to develop next-generation chemical risk assessment to advance research, share knowledge and improve skills, protecting human health and the environment. The present project is included in the PARC task focusing on facilitation of regulatory acceptance and use of new approach methodologies (NAMs).

A project group (PG) has been established with the responsibility for drafting the protocol and performing the study.

A scientific advisory group (SAG) consisting of experts in methods for tool development, systematic review methods, chemical risk assessment methods, toxicology, and/or NAMs has been established to share information about ongoing projects addressing similar questions to ensure that the outcome of this project complements the work of others and thereby creates synergies and avoids duplication of efforts, and to give strategic guidance and support.

2. Methods

2.1. General methodological issues

The performance testing consists of two exercises: the user testing, in which the application of INVITES-IN is evaluated in an emulated systematic review; and the creation of a "gold-standard" application of INVITES-IN, to which the results of the user testing can be compared. The gold-standard is a reference application of INVITES-IN. Cell culture studies having different degrees of bias will be identified through the process of the generation of the gold-standard and included in the practical exercise. This is important to ensure that the whole guidance document will be applied by the participants in the practical exercise. In addition, this should allow to interpret to what degree INVITES-IN succeeds to differentiate studies with a higher internal validity from studies with a lower internal validity.

2.1.1. Structure of the performance evaluation

An overview of the structure of the performance test of INVITES-IN is shown in Figure 1.

The creation of the gold-standard is described in Section 2.2.1. The user testing (Section 2.2.2) consists of a practical exercise (Section 2.2.2.2) and a survey (Section 2.2.2.3). The data collected will be used to assess consistency, assessor workload, accuracy, and user experience as shown in Figure 2.

An overview of the tasks and responsibilities of the PG and the SAG in the planning and execution of the user test and the creation of the gold-standard is given in Table 1.

An overview of the estimated workload for the participants is given in Figure 3. Participants not responding within the allocated deadline for completing the practical exercise will be excluded. Removed participants will not be replaced, unless it is necessary to reach the desired number of participants.

2.1.2. Ethical approval

Data Protection Impact Assessment has been performed and was approved by Norwegian Institute of Public Health on February 7th, 2022, (Archive No 22/04212). Ethical approval is not relevant since no biological materials or information on health will be collected.

2.2. Data collection

2.2.1. Selection of cell culture studies and creation of the gold-standard

The gold-standard will be created by a group of four to five PG members. All members in this group participate in the creation of gold-standard evaluations for each cell culture study included in the practical exercise. To harmonise the ratings given by the members in the group, all members start by evaluating the same three studies and then meet to compare and discuss the ratings. The members will then evaluate the same five studies and meet to discuss and compare the ratings, and so on. Consensus in the group will be required for completion of the gold-standards. The

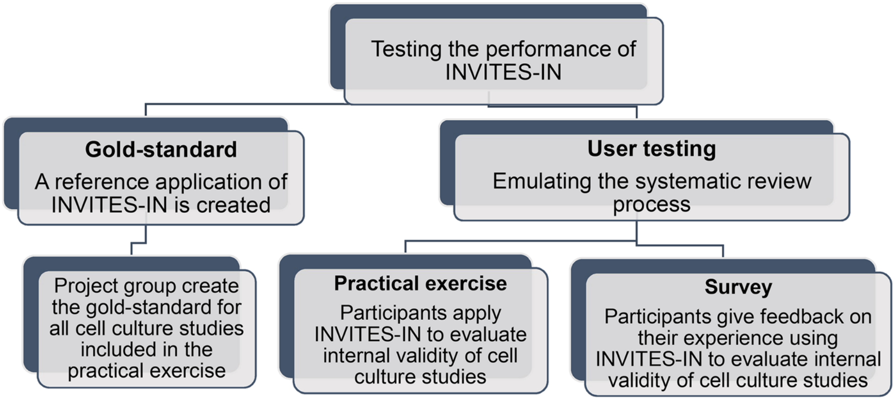
Testing the performance of INVITES-IN
Gold-standard
A reference application of INVITES-IN is created
User testing
Emulating the systematic review process
Project group create the gold-standard for all cell culture studies included in the practical exercise
Practical exercise
Participants apply INVITES-IN to evaluate internal validity of cell culture studies
Survey
Participants give feedback on their experience using INVITES-IN to evaluate internal validity of cell culture studies

Figure 1. The structure of the performance testing.

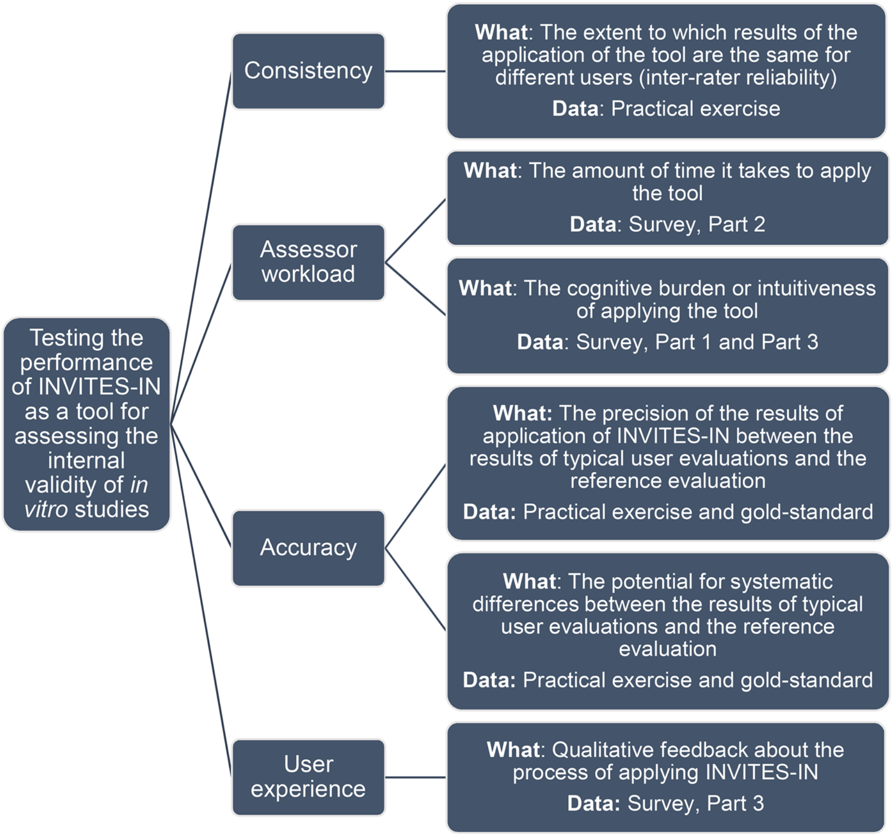
Testing the performance of INVITES-IN as a tool for assessing the internal validity of *in vitro* studies
Consistency
What: The extent to which results of the application of the tool are the same for different users (inter-rater reliability)
Data: Practical exercise
Assessor workload
What: The amount of time it takes to apply the tool
Data: Survey, Part 2
What: The cognitive burden or intuitiveness of applying the tool
Data: Survey, Part 1 and Part 3
Accuracy
What: The precision of the results of application of INVITES-IN between the results of typical user evaluations and the reference evaluation
Data: Practical exercise and gold-standard
What: The potential for systematic differences between the results of typical user evaluations and the reference evaluation
Data: Practical exercise and gold-standard
User experience
What: Qualitative feedback about the process of applying INVITES-IN
Data: Survey, Part 3

Figure 2. An overview of the performance testing.

gold-standard will be created during the process of identification of cell culture studies for the practical exercise as illustrated in Figure 4 (Step 5).

The procedure for identification of cell culture studies for the practical exercise and the creation of a gold-standard for each study includes seven steps (Figure 4). Systematic reviews including cell culture studies were identified in a literature search (step 1 in Figure 4). The eligibility criteria used for the selection of the systematic reviews were as follows:

a. A literature search was performed.
b. The population, exposure, comparator, outcome (PECO)/population, exposure, outcome (PEO) statement is clear.
c. Quality or RoB of the included studies has been assessed.
d. At least one of the included cell culture studies is categorised as high quality/low RoB and at least one is categorised as low quality/high RoB.

Criteria c) and d) were included in attempt to identify cell culture studies of varying RoB and varying quality, as all should be included in the user testing.

The literature search and study selection are described in Supplementary material 1, and includes search terms, search strategy, and study selection (Figure S1 (flow chart), Table S1 (included systematic reviews), and Table S2 (excluded publications, with reason)). An overview of the publications in the included systematic reviews that are categorised as high quality/low RoB or low quality/high RoB is given in Table S3 (step 2 in Figure 4). An overview of all cell culture studies included in these publications will be created (step 3 in Figure 4) and will be the starting point for the random selection of cell culture studies for the practical exercise (step 4 in Figure 4). The PG members developing INVITES-IN will create the gold-standard for these studies by rating all signalling question for each study according to the beta version of INVITES-IN (step 5 in Figure 4). The gold-standards should include more than one rating option for each signalling question (step 6 in Figure 4). If this is not achieved, steps 4 and 5 (Figure 4) will be repeated until this is fulfilled (step 7 in Figure 4). When this is fulfilled, these cell culture studies will constitute the pool of cell culture studies for the practical exercise.

Table 1. Tasks and responsibilities in the testing of INVITES-IN.

Phase	Task	Responsible
Planning	Define inclusion criteria for participants. Nominate and recruit participants fulfilling the inclusion criteria.	Project group and scientific advisory group
	Prepare the request for participant information. Select cell culture studies for the practical exercise and create the gold-standard. Prepare the survey.	Project group
Execution	The collection of participant information • Participants complete the questionnaire. The practical exercise • Pairs of testers are established. • Participants receive a cell culture study for training. • Training sessions are arranged. • Participants receive the cell culture studies. • Participants individually rate the cell culture studies. • Pair of participants (judge pairs) meet to compare, discuss, and reconcile their ratings. The survey • Participants complete the survey. Data analyses	Project group

The number of cell culture studies needed for the practical exercise depends on the number of user testing participants and this is addressed in Section 2.3.1.

2.2.2. User testing

2.2.2.1. Participant selection and enrolment. The participants in the user testing will be scientists having good working knowledge with *in vitro* methods, with or without systematic review and chemical risk assessment expertise, affiliated in academic institutions, governmental institutions (including risk assessment institutions and research institutes), and private sector research institutions (eligibility criteria are presented in Table 2).

The participants should be able to work in pairs during parts of the practical exercise (see Section 2.2.2.2). Participant information including affiliation, years of experience with *in vitro* studies, years of experience with chemical risk assessment, and years of experience with systematic reviews will be collected using a questionnaire (see Supplementary material 3).

No financial compensation or other incentives are offered for the participation; however, the participants will be eligible to be co-authors of the user testing study manuscript if they also read and comment on the final draft.

2.2.2.2. The practical exercise. The practical exercise aims to emulate a real-world application of INVITES-IN by simulating the validity assessment step in the systematic review process, albeit in a controlled environment. The validity assessment step includes critical appraisal of studies by pair of experts that first do the assessment individually and then harmonise the results.

Before the practical exercise is started, participants will receive one cell culture study for self-train on the application of INVITES-IN, and training sessions will be arranged to clarify initial questions and uncertainties. An overview of the most frequently asked questions will be created and made available for all participants.

If a publication contains several cell culture studies/experiments, a participant will only be asked to evaluate one of these. Participants will receive the publication and the selected cell culture experiment will be highlighted to ensure that INVITES-IN will be applied to the correct experiment.

To emulate the systematic review process, all cell culture studies included in the practical exercise will be tagged with a research question formulated based on the extracted PECO/PEO from the systematic review in which it was found (see 2.2.1 and Supplementary material 1). Each study will first be evaluated individually by a participant, followed by a calibration in pairs. The pairs will always be the same. When the practical exercise is started, the two participants in a pair will receive the same cell culture studies in a given order (the allocated studies for a pair will be numbered as study one, study two, etc.). Participants receive study one first and will not receive study two before they have sent the evaluation of study one to the PG. This way we ensure that the studies are evaluated in a given order.

Participants will be requested to time themselves, and to report the time used for the rating of each study in the survey they will receive when the practical exercise is completed (see Section 2.2.2.3). The process for the practical exercise is illustrated in Figure 5.

2.2.2.3. The survey. In the survey, consisting of three parts, participants will give feedback based on their experience using the beta version of INVITES-IN to evaluate internal validity (see Supplementary material 2).

In the first part, participants rate agreement with statements on the cognitive burden and intuitiveness of applying the tool. For each statement, participants may provide more details as free text. In the second part, participants are asked to time themselves when rating each cell culture study (cell culture study 1, cell culture study 2, etc.). They will also be asked how

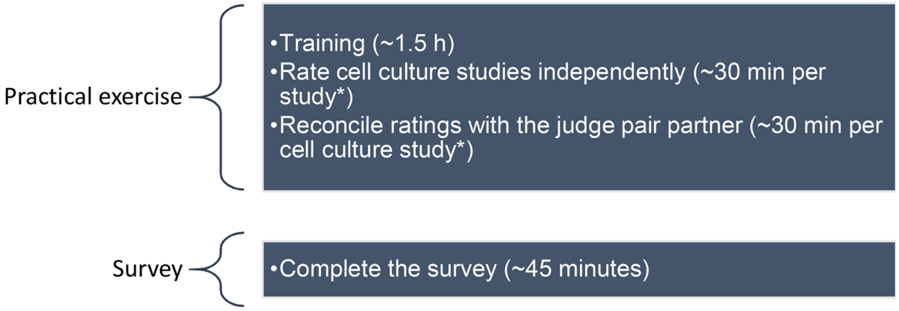
Practical exercise
•Training (~1.5 h)
•Rate cell culture studies independently (~30 min per study*)
•Reconcile ratings with the judge pair partner (~30 min per cell culture study*)
Survey
•Complete the survey (~45 minutes)

Figure 3. Estimated workload for the participants. *The number of cell culture studies each participant will receive depends on the total number of participants recruited (see Section 2.3.1).

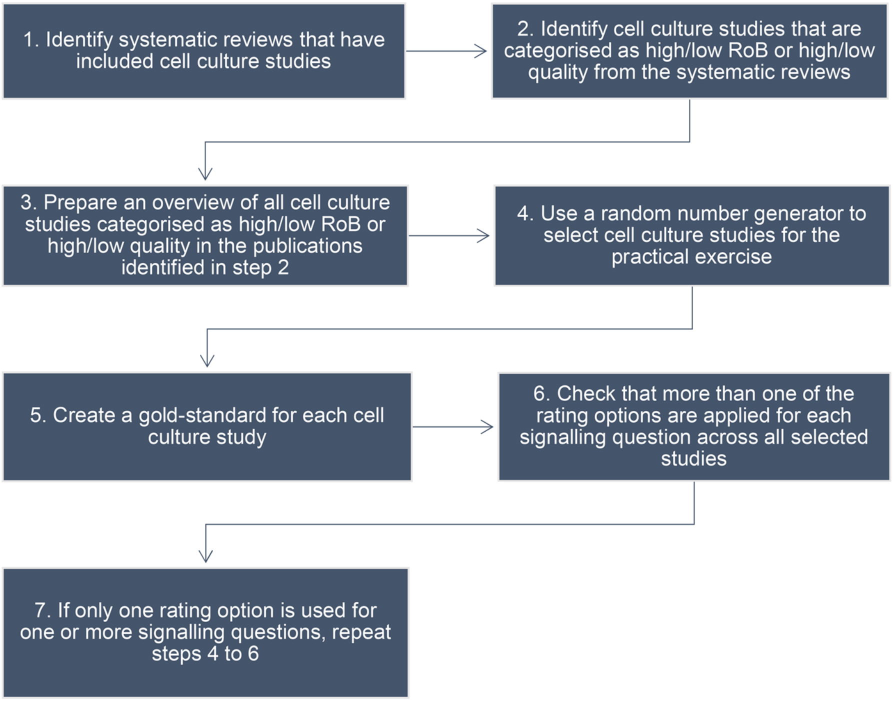
1. Identify systematic reviews that have included cell culture studies
2. Identify cell culture studies that are categorised as high/low RoB or high/low quality from the systematic reviews
3. Prepare an overview of all cell culture studies categorised as high/low RoB or high/low quality in the publications identified in step 2
4. Use a random number generator to select cell culture studies for the practical exercise
5. Create a gold-standard for each cell culture study
6. Check that more than one of the rating options are applied for each signalling question across all selected studies
7. If only one rating option is used for one or more signalling questions, repeat steps 4 to 6

Figure 4. The seven steps to create the pool of cell culture studies for the practical exercise and the gold-standard for each study.

Table 2. Eligibility criteria for participation in the user testing study.

Scientific experience or expertise	*In vitro* methods OR *In vitro* methods AND systematic review methods OR *In vitro* methods AND chemical risk assessment
Academic level	PhD degree completed
Language	English, level B1 or higher
Affiliation	Academia Governmental institutions (including risk assessment institutions and research institutes) Private sector research institutions NGOs Self-employed or freelancer researchers

confident they are that they timed themselves accurately, and given the options “very confident”, “somewhat confident”, “somewhat unsure”, or “very unsure”. In the third part, participants are requested to answer questions related to user experience,

2.3. Data analysis and reporting

All analyses will be done by the PG members.

All raw data will be anonymised and made available as supplementary to the user test report. The number of participants completing the practical exercise will be reported. All ratings of signalling questions, both pre- and post-reconciliation ratings, will be presented and made available. The number of participants completing the survey will be reported. All feedback on user experience will be made available.

Regarding participant information, participant response (in percentage), gender distribution (in percent), and experience with *in vitro* methods, chemical

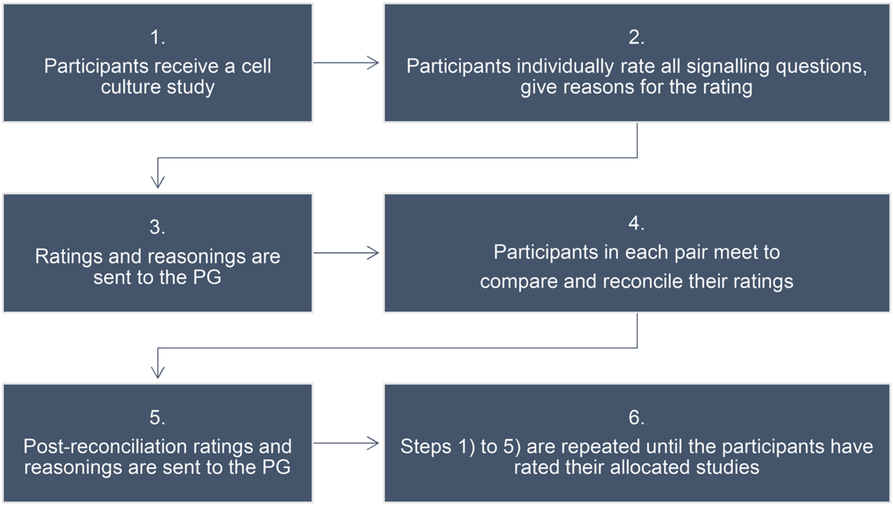
1.
Participants receive a cell culture study
2.
Participants individually rate all signalling questions, give reasons for the rating
3.
Ratings and reasonings are sent to the PG
4.
Participants in each pair meet to compare and reconcile their ratings
5.
Post-reconciliation ratings and reasonings are sent to the PG
6.
Steps 1) to 5) are repeated until the participants have rated their allocated studies

Figure 5. An overview of the practical exercise.

risk assessment and systematic review methods, will be reported.

2.3.1. Assessing consistency

Consistency in assessing internal validity using INVITES-IN, in terms of the extent to which results of the application of the tool are the same for different users, will be measured as inter-rater reliability using the intraclass correlation coefficient (ICC) of the studies. Low consistency may be due to varying level and specific field of expertise and/or insufficient or imprecise information in the guidance. The ICC calculations will be used to identify needs for improvement of the guidance. Data on participant expertise will be collected (see Supplementary materials 3) and used to see if there are correlations between the level of consistency and the level and type of expertise. However, statistical analyses addressing such effects will not be performed as we are not able to power the study for such analyses.

We will measure inter-rater reliability for the post-reconciliation ratings from the practical exercise (one rating per judge pair per cell culture study), and it will be measured per signalling question for all studies. The ratings for each signalling question will be ordinal, and the studies will be distributed in a weaved manner to counteract the bias from a learning. The assessment of consistency is described in detail in Section 2.3.1.1, and sample size estimation is described in detail in Section 2.3.1.2.

2.3.1.1. Detailed description of the statistical methods. The ratings for each signalling question will be ordinal with three, four, or five levels (Y = 1, 2, 3, Y = 1, 2, 3, 4, or Y = 1, 2, 3, 4, 5; the number of rating options will be decided during the creation of INVITES-IN). The ICC for each signalling question will be calculated for each signalling question via a mixed-effects linear regression model:

$$Y_{jk} = \mu + \alpha_j + \beta_k + \epsilon_{jk}$$

where Y_{jk} is the rating corresponding to study j that was rated by judge pair k. μ is the overall mean rating, α_j represents the study effect, β_k represents the judge pair effect, and ϵ_{jk} represents unexplained variation/random error effect.

The ICC (correlation of two observations of the same study) is then calculated by:

$$\text{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\epsilon^2}$$

where σ_α^2 is the variance of α (study effect), σ_β^2 is the variance of β (judge pair effect), and σ_ϵ^2 is the variance of ϵ (unexplained variation/random error). The total variation of the outcome (Y) can be expressed as $\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\epsilon^2$.

95% confidence intervals for the ICC are then obtained by performing model-based semi-parametric bootstrapping, using the lme4::bootMer function in R, and taking the 2.5th and 97.5th percentiles.

The ICC of the studies can be interpreted as the amount of variation (of the outcome) attributable to differences between the studies. The ICC value ranges from 0 to 1, with higher values indicating greater consistency of agreement between the ratings of a signalling question. That is, if the ICC of the studies is 0.6, the appropriate interpretation is that 60% of the total variance in the ratings is due to differences between the studies, and 40% of the total variance in the ratings is due to other causes (i.e. judge pair effect and random error).

The studies will be distributed in a weaved manner to counteract the bias from a learning effect as illustrated in Table 3. The resultant ICC can be interpreted roughly as the ICC after 3 attempts, as it is a blended ICC with equal number of studies rated in attempt 1, 2, 3, 4, and 5.

2.3.1.2. Sample size estimation. The number of studies that will be rated by a pair depends on the number of participants that we manage to recruit (Tables 4–6). Sample size estimates were made through simulations.

The simulation process for four possible ratings of each signalling question ($Y=1, 2, 3, 4$; Table 5) is described in detail. The scenarios where the possible ratings of each signalling question are three ($Y=1, 2, 3$; Table 4) or five ($Y=1, 2, 3, 4, 5$; Table 6) were simulated in a similar manner and hence do not need to be described. A total of 100 000 fictional studies were simulated, of which 25 000 corresponded to $Y^{\text{true}}=1$, 25 000 corresponded to $Y^{\text{true}}=2$, and so on.

From here, multiple scenarios were considered, where the number of judge pairs and the number of studies reviewed per judge pair were altered. For each scenario, the 100 000 studies (or less, as appropriate) were assigned to 10 000 judge pairs. A new outcome (Y) was created to provide an ICC of approximately 0.65. For 78.5% of the judge pair/paper combinations, Y was equal to Y^{true} (i.e. the judge pair accurately assessed the paper). To create disagreements between the judges, for 21.5% of the judge pair/paper combinations, a number was randomly selected between 1 and 4 (R code provided in supplemental materials 4).

$$\Pr\left(Y_{jk}=Y_{jk}^{true}\right)=0.785$$

$$\Pr\left(Y_{jk}=random\left(1,2,3,4\right)\right)=0.215$$

where Y_{jk} is the rating corresponding to study j that was rated by judge pair k, as previously noted in Section 2.3.1.1.

The ICC estimation was then performed multiple times (e.g. if the scenario needed 10 judge pairs, then the ICC estimation was performed 1000 times) as explained in Section 2.3.1.1.

For each of the ICC estimations, the margin of error (MOE) was calculated (MOE = half the width of the bootstrapped 95% confidence interval). The mean MOE of each scenario is presented in Tables 4–6. The mean MOE was chosen to be presented as it represents how reliable our estimates are (R code provided in Supplemental materials 4).

A maximum MOE of 0.15 was considered to be acceptable. According to Koo and Li (2016), ICC values less than 0.5 are indicative of poor inter-rater reliability for a signalling question, whereas ICC values between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of moderate, good, and excellent reliability, respectively. By estimating the MOE when the true ICC is 0.65 and setting the acceptable MOE to 0.15, we will be able to differentiate between poorer (ICC less than 0.5) and better (ICC ≤ 0.5) inter-rater reliability.

2.3.2. Assessing assessor workload

Assessor workload in applying INVITES-IN will be assessed as the amount of time it takes to apply the tool, and the cognitive burden or intuitiveness of applying the tool. The participants will be instructed to time all time spent on the evaluation, which includes reading the study, identifying the necessary information, conducting additional information gathering (when necessary) and answering and giving reasoning for all the signalling questions.

Median and mean time to complete the assessments of one cell culture study (part 2 of the survey) will be reported. Subgroup analyses may be performed to evaluate the effect of having systematic review experience.

Table 3. Illustrating the distribution of studies to the judge pairs.

	Judge pair (40 participants = 20 judge pairs)																			
Study	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	X1	P					X5	X4	X3	X2	X1	P					X5	X4	X3	X2
2	X2	X1	P					X5	X4	X3	X2	X1	P					X5	X4	X3
3	X3	X2	X1	P					X5	X4	X3	X2	X1	P					X5	X4
4	X4	X3	X2	X1	P					X5	X4	X3	X2	X1	P					X5
5	X5	X4	X3	X2	X1	P					X5	X4	X3	X2	X1	P				
6		X5	X4	X3	X2	X1	P					X5	X4	X3	X2	X1	P			
7			X5	X4	X3	X2	X1	P					X5	X4	X3	X2	X1	P		
8				X5	X4	X3	X2	X1	P					X5	X4	X3	X2	X1	P	
9					X5	X4	X3	X2	X1	P					X5	X4	X3	X2	X1	P
10	P					X5	X4	X3	X2	X1	P					X5	X4	X3	X2	X1

A red P demarcates a practice study, which will not be included in the final analysis. X1 demarcates the first study and the judge pair's first attempt to apply INVITES-IN, X2 demarcates the second study and the judge pair's second attempt to apply INVITES-IN, etc.

Table 4. Sample size estimates for the number of cell culture studies needed to be assessed by the judge pairs to have a MOE ≤ 0.15 given three rating options of a signalling question (Y = 1, 2, 3).

95% margin of error for estimating ICC when the true ICC is 0.65, for Y = 1, 2, 3						
Judge pairs	Number of cell culture studies reviewed per judge pair					
	5	6	7	8	9	10
10	0.23	0.21	0.20	0.18	0.18	0.17
12	0.21	0.20	0.19	0.17	0.17	0.16
14	0.20	0.18	0.17	0.16	**0.15**	**0.14**
16	0.19	0.17	0.16	**0.15**	**0.15**	**0.14**
18	0.18	0.16	0.16	**0.14**	**0.14**	**0.13**
20	0.16	**0.15**	**0.14**	**0.13**	**0.13**	**0.12**
22	0.16	**0.15**	**0.14**	**0.13**	**0.12**	**0.12**
24	0.16	**0.14**	**0.14**	**0.13**	**0.12**	**0.12**
26	**0.15**	**0.14**	**0.13**	**0.12**	**0.12**	**0.11**
28	**0.14**	**0.13**	**0.12**	**0.11**	**0.11**	**0.10**
30	**0.14**	**0.13**	**0.11**	**0.11**	**0.11**	**0.10**

ICC: intraclass correlation coefficient; MOE: margin of error; Y: rating option.

Table 5. Sample size estimates for the number of cell culture studies needed to be assessed by the judge pairs to have a MOE ≤ 0.15 given four rating options of a signalling question (Y = 1, 2, 3, 4).

95% margin of error for estimating ICC when the true ICC is 0.65, for Y = 1, 2, 3, 4						
Judge pairs	Number of cell culture studies reviewed per judge pair					
	5	6	7	8	9	10
10	0.23	0.23	0.21	0.20	0.19	0.17
12	0.21	0.21	0.20	0.18	0.17	**0.15**
14	0.21	0.19	0.19	0.18	0.17	**0.15**
16	0.19	0.18	0.17	0.16	**0.15**	**0.14**
18	0.18	0.17	0.16	**0.15**	**0.14**	**0.13**
20	0.17	0.16	**0.15**	**0.14**	**0.13**	**0.12**
22	0.17	0.16	**0.15**	**0.14**	**0.13**	**0.12**
24	0.16	**0.15**	**0.15**	**0.13**	**0.13**	**0.11**
26	**0.15**	**0.14**	**0.14**	**0.13**	**0.12**	**0.11**
28	**0.15**	**0.14**	**0.13**	**0.12**	**0.12**	**0.10**
30	**0.15**	**0.13**	**0.13**	**0.12**	**0.11**	**0.10**

ICC: intraclass correlation coefficient; MOE: margin of error; Y: rating option.

Table 6. Sample size estimates for the number of cell culture studies needed to be assessed by the judge pairs to have a MOE ≤ 0.15 given five rating options of a signalling question (Y = 1, 2, 3, 4, 5).

95% margin of error for estimating ICC when the true ICC is 0.5, for Y = 1, 2, 3, 4, 5						
Judge pairs	Number of cell culture studies reviewed per judge pair					
	5	6	7	8	9	10
10	0.23	0.23	0.19	0.19	0.16	0.17
12	0.21	0.21	0.17	0.18	**0.15**	**0.15**
14	0.20	0.20	0.17	0.16	**0.14**	**0.15**
16	0.18	0.18	**0.15**	**0.15**	**0.13**	**0.13**
18	0.18	0.17	**0.15**	**0.14**	**0.12**	**0.13**
20	0.17	0.16	**0.14**	**0.14**	**0.12**	**0.12**
22	0.16	0.16	**0.14**	**0.14**	**0.11**	**0.12**
24	0.16	**0.15**	**0.13**	**0.13**	**0.11**	**0.11**
26	**0.15**	**0.15**	**0.12**	**0.12**	**0.10**	**0.11**
28	**0.14**	**0.14**	**0.12**	**0.12**	**0.10**	**0.11**
30	**0.14**	**0.13**	**0.11**	**0.12**	**0.10**	**0.10**

ICC: intraclass correlation coefficient; MOE: margin of error; Y: rating option.

All data on the cognitive burden or intuitiveness (part 1 of the survey) of applying the tool will be reported.

2.3.3. Assessing accuracy

Accuracy of the tool, in terms of precision of the results of application of INVITES-IN and the potential for systematic differences between the results of typical user evaluations of a study and the gold-standard reference evaluation, will be measured by comparing the post-reconciliation ratings and the gold-standard. How precisely INVITES-IN characterises internal validity will be measured as the association between the ratings of participants and the gold-standard for the same signalling question for the same study and by identification of learning effects (Section 2.3.3.1). The potential for systematic differences will be described via histograms of the user-ratings for each level of the gold-standard ratings (Section 2.3.3.2).

2.3.3.1. Precision. To see if there is an association between the gold-standard reference rating and the judge pair ratings, a matrix of ratings will be formed with the gold-standard forming one axis and the post-reconciliation ratings forming the other axis. From this matrix, two analyses will be performed. Firstly, Fisher's exact test will be performed to assess if there is evidence that the user-ratings are statistically associated with the gold-standard ratings. Results indicating no difference between the gold-standard and the user-ratings are shown in Table 7.

The secondary analysis will be primarily used to attempt to identify if there is a learning effect. The aforementioned analyses (Table 7) will be repeated five times, with data restricted to the judge pair's first attempt to apply INVITES-IN to evaluate the internal validity of a cell culture study, second attempt, third attempt, fourth attempt, and fifth attempt, meaning that a study will be assessed five times. Furthermore, matrix of ratings will be constructed where the data is restricted to gold-standard Y = 1, Y = 2, Y = 3, Y = 4, one axis corresponds to first attempt, second attempt, third attempt, fourth attempt, and the fifth attempt to apply INVITES-IN to a cell culture study, and the other axis corresponds to the post-reconciliation ratings (Y = 1, 2, 3, 4) (Tables 8–10). Results indicating a small difference between gold-standard and user-ratings, no difference between gold-standard and user-ratings, and a learning effect are shown in Tables 8–10, respectively.

2.3.3.2. Bias. The potential for systematic difference between user-ratings and the gold-standard will be described via histograms of the user-ratings for each

Table 7. For all data from the practical exercise; illustrating results indicating no difference between the gold-standard and user-ratings.

		Gold-standard rating			
All data (SQ = 1)		Y = 1	Y = 2	Y = 3	Y = 4
Judge pair ratings	Number of judge pairs selecting the rating option Y = 1	10	0	0	0
	Number of judge pairs selecting the rating option Y = 2	0	10	0	0
	Number of judge pairs selecting the rating option Y = 3	0	0	10	0
	Number of judge pairs selecting the rating option Y = 4	0	0	0	10

Comparison of the gold-standard with ratings from 40 study assessments, where the ratings are equally divided for the different rating options (Y = 1, 2, 3, or 4). The gold-standard rating is indicated in the table with grey colouring of table cells. This analysis will be repeated for all signalling questions.
SQ: signalling question; Y: rating option.

Table 8. Data restricted to the first attempt to apply INVITES-IN to a cell culture study; results indicating small difference between gold-standard and user-ratings.

		Gold-standard rating			
Restrict data to first attempt (SQ = 1)		Y = 1	Y = 2	Y = 3	Y = 4
Judge pair ratings	Number of judge pairs selecting the rating option Y = 1	7	1	1	1
	Number of judge pairs selecting the rating option Y = 2	1	7	1	1
	Number of judge pairs selecting the rating option Y = 3	1	1	7	1
	Number of judge pairs selecting the rating option Y = 4	1	1	1	7

Comparison of the gold-standard with ratings from 40 study assessments, where the rating options are Y = 1, 2, 3, or 4. The gold-standard rating is indicated in the table with grey colouring of table cells. This analysis will be repeated for all signalling questions.
SQ: signalling question; Y: rating option.

Table 9. Data restricted to the second attempt to apply INVITES-IN to a cell culture study; results indicating no difference between gold-standard and user-ratings.

		Gold-standard rating			
	Restrict data to second attempt (SQ = 1)	Y = 1	Y = 2	Y = 3	Y = 4
Judge pair ratings	Number of judge pairs selecting the rating option Y = 1	10	0	0	0
	Number of judge pairs selecting the rating option Y = 2	0	10	0	0
	Number of judge pairs selecting the rating option Y = 3	0	0	10	0
	Number of judge pairs selecting the rating option Y = 4	0	0	0	10

Comparison of the gold-standard with ratings from 40 study assessments, where the rating options are Y = 1, 2, 3, or 4. The gold-standard rating is indicated in the table with grey colouring of table cells. This analysis will be repeated for all signalling questions.
SQ: signalling question; Y: rating option.

Table 10. Results indicating a learning effect, assuming 40 judge pairs rating one signalling question for 40 studies with a gold-standard rating Y = 1.

Restrict data to gold-standard Y = 1 (SQ = 1)	Attempt 1	Attempt 2	Attempt 3	Attempt 4	Attempt 5
Number of judge pairs selecting the rating option Y = 1	10	20	25	30	35
Number of judge pairs selecting the rating option Y = 2	10	10	15	10	5
Number of judge pairs selecting the rating option Y = 3	10	10	0	0	0
Number of judge pairs selecting the rating option Y = 4	10	0	0	0	0

The rating options are Y = 1, 2, 3, or 4. The data is restricted to when the gold-standard rating is Y = 1. This analysis will be repeated for all signalling questions, and for Y = 1, 2, 3, and 4.
SQ: signalling question; Y: rating option.

level of the gold-standard ratings due to the small number of categories in the outcome (i.e. one histogram of user-ratings when gold-standard rating = 1, another histogram of user-ratings when gold-standard rating = 2, …). If the Fisher's exact test (Section 2.3.3.1) is statistically significant, then these histograms (along with additional summary statistics, such as the mean/median, or cumulative distribution function) will be interpreted to identify if the user-ratings are positively associated with the gold-standard reference ratings.

2.3.4. Reporting qualitative data on user experience

The user experience in applying the tool (Part 3 of the survey), in terms of qualitative feedback about the process of applying INVITES-IN, will constitute of free text answers to questions. Overviews of all answers to each question will be prepared and made available. If possible, we will identify trends in the answers to each question, and the answers will then be grouped according to these trends. Overviews of the rating of each question will be prepared and made available.

3. Results and creation of the release version of INVITES-IN

All results from the data analyses will be used to evaluate the performance of INVITES-IN as a tool for assessing the internal validity of *in vitro* studies. Creating the release version of INVITES-IN, these results will be the basis for all discussions regarding needs for adjustments of the INVITES-IN version that was applied for the user testing.

Results from the practical exercise, such as if data on reasoning for the rating of a signalling question clearly shows that several of the test users have interpreted the guidance incorrectly or differently, will be used to identify needs for improvement of the guidance together with the feedback from the survey. In addition, poor inter-rater reliability will be interpretated as needs for improvement/revision such as clarification of the guidance.

The structure of INVITES-IN, layout, and other presentational or user-experience related characteristics can be adjusted in response to the outcome of user testing, e.g. in response to feedback on the survey questions addressing how the tool, and the usability of the tool, can be improved.

The signalling questions will not be revised, unless in the judgement of the PG, informed by the SAG, there are sufficient grounds for changing a signalling question, and there is good reason for believing the change made will improve INVITES-IN. This may be done in response to feedback on the survey question addressing which signalling questions the users found most challenging to answer, and why this question was particularly challenging.

The PG and SAG will decide on which improvements should be performed to the guidance and the structure/layout, and justification for the decisions will be included in the user testing report. PG members involved in this study will make the final decisions on improvements and prepare the release version.

4. Limitations

By recruiting from the network of PG and SAG members, we attempt to involve a broad and diverse group of experts from different institutions in the user testing, while working within the resources that the research team has available. However, the final group of participants may not be representative of all potential users of the tool."

The study is not powered to allow meaningful analysis of impact of participant characteristics on the results. It was considered not possible to power the study for such analyses because the number of participants needed would be higher than can be achieved with the resources available to the research team.

The study is not powered to allow analysis of impact of *in vitro* studies on the results. It was considered not possible to power the study for such analyses because the number of studies that would need analysing is unrealistically high.

Dissemination

A report describing the user testing study, including all results, will be published.

The final version of INVITES-IN will be published as peer-review journal article and disseminated through PG and SAG members.

Definitions

Bias is systematic errors, or deviations from the truth, in results or inference (Cochrane Collaboration 2005).

Bias domains are themes such as e.g. study performance, analysis, and reporting, under which sources of bias/bias items can be organised/grouped.

Bias items are study properties that may be relevant for introduction of bias in results and/or their interpretation. Criteria are the issues that have to be fulfilled for bias to be avoided. In the guidance document for the INVITES-IN tool there will be criteria for reaching risk-of-bias judgements for each signalling question.

Cell culture studies refer to cell lines, primary cells, co-cultures, monolayer, and 3-D cell systems in the user testing study.

Internal validity is the extent to which the design and conduct of a study are likely to have prevented bias (Cochrane Collaboration 2005).

Inter-rater reliability refers to the level of consistency between independent raters when rating the same signalling question for the same study.

In vitro ("in the glass") tests mean that it is done outside of a living organism and it usually involves isolated tissues, organs or cells (ECHA 2023).

NAMs have not yet a standard definition. However, there seems to be a general agreement that the term "NAMs" include *in chemico*, *in silico* and *in vitro* studies. One definition is that "NAMs includes any technology, methodology, approach, or combination that can provide information on chemical hazard and risk assessment without the use of animals, including *in silico, in chemico, in vitro*, and *ex vivo* approaches" (ECHA 2016; EPA 2018).

Risk of bias is a measure for systematic errors. Risk of bias tools is used for evaluation of the extent to which the design and conduct of a study are likely to have prevented bias (the degree of systematic errors).

Signalling questions are the questions that the users of the tool answer in order to determine whether the criteria have been fulfilled.

Validity is the degree to which a result (of a measurement or study) is likely to be true and free of bias (systematic errors) (Cochrane Collaboration 2005).

Ethical considerations

Data Protection Impact Assessment has been performed and was approved by Norwegian Institute of Public Health on February 7th, 2022 (Archive No 22/04212). Ethical approval is not relevant since no biological materials or information on health will be collected.

Author contribution

Conceptualization: Gro H. Mathisen, Gunn E. Vist, Paul Whaley, Richard A. White, Trine Husøy, and Camilla Svendsen.

Data curation: Gro H. Mathisen, Gunn E. Vist, Paul Whaley, Richard A. White, and Camilla Svendsen.

Funding acquisition: Gro H. Mathisen and Camilla Svendsen.

Methodology: Gro H. Mathisen, Gunn E. Vist, Paul Whaley, Richard A. White, Trine Husøy, and Camilla Svendsen.

Project administration: Gro H. Mathisen and Camilla Svendsen.

Supervision: Paul Whaley, Anna Beronius, Ingrid Druwe, Thomas Hartung, Sebastian Hoffmann, Carlijn R. Hooijmans, Pilar Prieto-Peraita, Joshua F. Robinson, Erwin Roggen, Andrew A. Rooney, Mathieu Vinken.

Visualization: Gro H. Mathisen.

Writing – original draft: Gro H. Mathisen, Gunn E. Vist, Paul Whaley, Richard A. White, and Camilla Svendsen.

Writing – review & editing: Gro H. Mathisen, Gunn E. Vist, Paul Whaley, Richard A. White, Trine Husøy, Heather M. Ames, Anna Beronius, Emma Di Consiglio, Ingrid Druwe, Thomas Hartung, Sebastian Hoffmann, Carlijn R. Hooijmans, Kyriaki Machera, Pilar Prieto, Joshua F. Robinson, Erwin Roggen, Andrew A. Rooney, Nicolas Roth, Eliana Spilioti, Anastasia Spyropoulou, Olga Tcheremenskaia, Emanuela Testai, Mathieu Vinken, and Camilla Svendsen.

Disclosure statement

Completed declaration of interest forms for each author is available as supplementary materials. The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this protocol.

Disclaimers

The author Ingrid Druwe is employed at the U.S. Environmental Protection Agency. The views expressed in this manuscript are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

The author Andrew Rooney is employed at the U.S. National Institute of Environmental Health Sciences. The views expressed in this manuscript are those of the authors and do not necessarily represent the views or policies of the U.S. National Institute of Environmental Health Sciences.

Funding

Gro H. Mathisen, Gunn E. Vist, Paul Whaley, Richard A. White, Trine Husøy, Heather M. Ames, Emma Di Consiglio, Kyriaki Machera, Eliana Spilioti, Anastasia Spyropoulou, Olga Tcheremenskaia, Emanuela Testai, and Camilla Svendsen were supported by HORIZON-HLTH-2021-ENVHLTH-03, Grant [101057014]; Nicolas Roth was supported by The Swiss State Secretariat for Education, Research and Innovation (SERI) under contract no. 22.00230.

ORCID

Gro H. Mathisen http://orcid.org/0000-0003-3824-4898
Paul Whaley http://orcid.org/0000-0003-4021-0785
Richard A. White http://orcid.org/0000-0002-6747-1726
Trine Husøy http://orcid.org/0000-0002-2827-7111
Heather M. Ames http://orcid.org/0000-0001-8509-7160
Anna Beronius http://orcid.org/0000-0001-9494-5395
Emma Di Consiglio http://orcid.org/0000-0001-5575-0376
Ingrid Druwe http://orcid.org/0000-0002-0591-0295
Thomas Hartung http://orcid.org/0000-0003-1359-7689
Sebastian Hoffmann http://orcid.org/0000-0002-3214-7678
Carlijn R. Hooijmans http://orcid.org/0000-0001-6435-5714
Kyriaki Machera http://orcid.org/0000-0001-6499-3468
Pilar Prieto http://orcid.org/0000-0002-9275-6248
Joshua F. Robinson http://orcid.org/0000-0002-2421-4535
Erwin Roggen http://orcid.org/0000-0002-9371-8390
Andrew A. Rooney http://orcid.org/0000-0002-1756-5185
Nicolas Roth http://orcid.org/0000-0002-2917-4384
Eliana Spilioti http://orcid.org/0000-0002-0441-7466
Anastasia Spyropoulou http://orcid.org/0000-0001-8599-7740
Olga Tcheremenskaia http://orcid.org/0000-0002-5029-3484
Emanuela Testai http://orcid.org/0000-0003-3113-5103
Mathieu Vinken http://orcid.org/0000-0001-5115-8893
Camilla Svendsen http://orcid.org/0000-0002-3216-0469

References

Cochrane Collaboration. 2005. Glossary of Terms in The Cochrane Collaboration, Version 4.2.5. http://aaz.hr/resources/pages/57/7.%20Cochrane%20glossary.pdf.

ECHA. 2016. "New Approach Methodologies in Regulatory Science: Proceedings of a Scientific Workshop: Helsinki," 1–14 April 2016, European Chemicals Agency. https://echa.europa.eu/documents/10162/21838212/scientific_ws_proceedings_en.pdf/a2087434-0407-4705-9057-95d9c2c2cc57.

ECHA. 2023. "In Vitro Methods", European Chemicals Agency. https://echa.europa.eu/support/registration/how-to-avoid-unnecessary-testing-on-animals/in-vitro-methods.

EPA. 2018. «Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the

TSCA Program. U.S. Environmental Protection Agency". EPA-740-R1-8004, United States Environmental Protection Agency. https://www.epa.gov/sites/default/files/2018-06/documents/epa_alt_strat_plan_6-20-18_clean_final.pdf.

Koo, T. K., and M. Y. Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15 (2): 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Moher, D., A. Liberati, J. Tetzlaff, and D. G. Altman. 2010. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement." *International Journal of Surgery (London, England)* 8 (5): 336–341. https://doi.org/10.1016/j.ijsu.2010.02.007

Svendsen, C., P. Whaley, G. E. Vist, T. Husøy, A. Beronius, E. Di Consiglio, I. Druwe, et al. 2023. "Protocol for Designing INVITES-IN, a Tool for Assessing the Internal Validity of in Vitro Studies." *Evidence-Based Toxicology* 1 (1): 2232415. https://doi.org/10.1080/2833373X.2023.2232415

VKM. 2023. "The Norwegian Scientific Committee of Food and Environment (VKM) Participates in the European Partnership for the Assessment of Risks from Chemicals (PARC)", Assessed May 11, 2023. https://vkm.no/english/parc/parceuropeanpartnershipfortheassessmentofrisksfromchemicals.4.7205492a1864a8c8da2dcfd9.html.

Whiting, P. F., R. Wolff, S. Mallett, I. Simera, and J. Savović. 2017. "A Proposed Framework for Developing Quality Assessment Tools." *Systematic Reviews* 6 (1): 204. https://doi.org/10.1186/s13643-017-0604-6