1   Protocol for designing INVITES-IN, a tool for assessing the internal validity of *in vitro* studies

2   Camilla Svendsen[1,2], Paul Whaley[1,3], Gunn E. Vist[1,4], Trine Husøy[1,5], Anna Beronius[6], Emma

3   Di Consiglio[7], Ingrid Druwe[8], Thomas Hartung[9,10], Vasiliki I. Hatzi[11], Sebastian Hoffmann[12,13],

4   Carlijn R Hooijmans[14], Georges Kass[15], Kyriaki Machera[11], Joshua F. Robinson[16], Erwin

5   Roggen[17], Andrew A. Rooney[18], Nicolas Roth[19,20], Eliana Spilioti[11], Anastasia Spyropoulou[11],

6   Olga Tcheremenskaia[7], Emanuela Testai[7], Mathieu Vinken[21], Gro H. Mathisen[1†]

7   [1]Norwegian Scientific Committee for Food and Environment, Norwegian Institute of Public

8   Health, Oslo, Norway

9   [2]Department of Chemical Toxicology, Norwegian Institute of Public Health, Oslo, Norway

10   [3]Lancaster Environment Centre, Lancaster University, Lancaster, UK

11   [4]Division for Health Services, Norwegian Institute of Public Health, Oslo, Norway

12   [5]Department of Food Safety, Norwegian Institute of Public Health, Oslo, Norway

13   [6]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

14   [7]Environment & Health Department, Italian National Institute of Health (ISS), Rome, Italy

15   [8]United States Environmental Protection Agency, Office of Research and Development,

16   Center for Public Health and Environmental Assessments, Research Triangle Park, NC,

17   USA.

18   [9]Center for Alternatives to Animal Testing (CAAT), Johns Hopkins University, Bloomberg

19   School of Public Health, Baltimore, MD, USA

20   [10]CAAT Europe, University of Konstanz, Konstanz, Germany

21   [11]Laboratory of Toxicological Control of Pesticides, Scientific Directorate of Pesticides'

22   Control and Phytopharmacy, Benaki Phytopathological Institute, Kifissia, Greece

23   [12]Evidence-Based Toxicology Collaboration (EBTC), Johns Hopkins University, Bloomberg

24   School of Public Health, Baltimore, MD, USA

25   [13]seh consulting + services, Paderborn, Germany

26   [14]Department of Anesthesiology, Pain and Palliative Care, Radboud University Medical

27   Centre, Nijmegen, Netherlands

28   [15]European Food Safety Authority, Parma, Italy

29 [16]Department of Obstetrics, Gynecology & Reproductive Sciences, University of California,

30 San Francisco (UCSF), USA

31 [17]3Rs Management and Consulting ApS, Lyngby, Denmark

32 [18]Division of Translational Toxicology, National Institute of Environmental Health Sciences,

33 Research Triangle Park, NC, USA.

34 [19]Department of Pharmaceutical Sciences, University of Basel, Basel, Switzerland

35 [20]Swiss Centre for Applied Human Toxicology (SCAHT)

36 [21]Department of Pharmaceutical and Pharmacological Sciences, Vrije Universiteit Brussel;

37 Belgium

38 [†]Correspondence should be addressed to Gro H. Mathisen; E-mail:

39 gro.haarklou.mathisen@vkm.no

40 Disclaimer

41 The author Ingrid Druwe is employed at the U.S. Environmental Protection Agency. The

42 views expressed in this manuscript are those of the authors and do not necessarily

43 represent the views or policies of the U.S. Environmental Protection Agency.

44 The author George Kass is employed at the European Food Safety Authority. The views and

45 opinions expressed in this manuscript are those of the contributing authors and do not

46 necessarily represent those of EFSA.

47 The author Andrew Rooney is employed at the U.S. National Institute of Environmental

48 Health Sciences. The views expressed in this manuscript are those of the authors and do

49 not necessarily represent the views or policies of the U.S. National Institute of Environmental

50 Health Sciences.

51 Abstract

52 This protocol describes the design and development of a tool for evaluation of the internal

53 validity of *in vitro* studies, which is needed to include the data as evidence in systematic

54 reviews and chemical risk assessments. The tool will be designed specifically to be applied

55 to cell culture studies, including, but not restricted to studies meeting the new approach

56 methodology (NAM) definition. The tool is called INVITES-IN (IN VITro Experimental Studies

57 INternal validity). Methods to generate evidence for regulatory toxicology are increasingly

58 shifting from classical animal experiments to new approach methodologies (NAMs), with a

59 corresponding need for methods to incorporate them in systematic reviews and other

60 evidence synthesis processes. While many tools have been created for assessing *in vitro*

61 ~~studies, no single tool is obviously authoritative specifically for assessing the internal validity~~
62 ~~of *in vitro* study designs. We therefore aim to address this situation by developing a new tool,~~
63 ~~INVITES-IN, for evaluating the internal validity of *in vitro* studies, using methods that ensure~~
64 ~~we are building on prior work, with a degree of rigour consistent with our intent to provide an~~
65 ~~authoritative assessment tool.~~

66 In this protocol, ~~The current protocol describes~~ three of the four studies that will be
67 performed to create the release version of INVITES-IN are described. In the first study,
68 evaluation of existing assessment tools will be combined with focus group discussions to
69 identify how characteristics of the design or conduct of an *in vitro* study can affect its internal
70 validity. Bias domains and items considered to be of relevance for *in vitro* studies will be
71 identified. In the second study, group agreement on internal validity domains and items of
72 importance for *in vitro* studies will be identified via a modified Delphi methodology. In the
73 third study, the draft version of the tool~~INVITES-IN~~ will be created, based on the data on
74 relevance and importance of bias domains and items collected in studies one and two. A
75 separate protocol will be prepared for the fourth study, which includes the user testing and~~,~~
76 validation of the tool, and collection of users' experience.

Key words
78 Cell culture, NAMs, Next Generation Risk Assessment, risk of bias.

1. Introduction
1.1 Evaluation of internal validity

81 This protocol describes the design and development of a tool for evaluation of the internal
82 validity of *in vitro* studies. ~~The tool is called INVITES-IN (IN VITro Experimental Studies~~
83 ~~INternal validity).~~ Internal validity is the extent to which a study (methodological design,
84 methods, and data analysis) is free from bias, where bias is "systematic error, or deviation
85 from the truth, in results" (Cochrane Collaboration, 2005). A test performed *in vitro* ("in the
86 glass") means that it is done outside of a living organism and it usually involves isolated
87 tissues, organs or cells (ECHA, 2023). The tool is called INVITES-IN (IN VITro Experimental
88 Studies INternal validity).

89 Methods to generate evidence for regulatory toxicology are shifting from classical animal
90 experiments to new approach methodologies (NAMs). The European Chemicals Agency and
91 the U.S. Environmental Protection Agency define NAMs as any technology, methodology,
92 approach, or combination that can provide information on chemical hazard and risk
93 assessment without the use of animals, including *in silico*, *in chemico*, *in vitro*, and *ex vivo*
94 approaches (ECHA, 2016; EPA, 2018). According to the European Food Safety Authority

95 (EFSA) the term NAMs is used to make reference to any non-animal-based approach that

96 can be used to provide toxicological information in the context of hazard/risk assessments

97 (EFSA et al., 2022). ~~Although no standard definition of NAMs is currently agreed upon, there~~

98 ~~seems to be a general agreement that the term "NAMs" includes *in chemico*, *in silico* and *in*~~

99 ~~*vitro* studies.~~

100

101 As part of the gradual incorporation and transition toward the use of NAMs, including *in vitro*

102 *studies,* a framework for evidence-based use of NAMs in toxicological research and

103 chemical risk assessment is required. Such a framework should ultimately incorporate at

104 least the following principles:

1. Result in identification of all relevant NAM-generated evidence relating to the
   research question addressed in a systematic review or risk assessment.
2. Provide for the evaluation of the internal validity of NAM studies (propensity for
   systematic error due to how the study is designed and conducted).
3. Provide for the evaluation of the external validity of NAM studies (the degree to which
   results of a study can be translated/generalised to human adverse health effects).
4. Contribute to objectivity, robustness, transparency, and reproducibility in the hazard
   identification and characterisation process.
5. In its approach to normalising and structuring the description and analysis of NAMs,
   contribute to progress in the extent to which research data conforms to FAIR
   (Findable, Accessible, Interoperable and Re-usable) principles of open science.

116 Systematic review and evidence-based toxicology principles should be implemented in all

117 parts of the framework, and it should be generic and usable across different regulatory

118 sectors such as food safety, cosmetic ingredient safety, etc. Principles for incorporating

119 evidence from NAMs into risk assessments and a framework for the evaluation of skin

120 sensitisation have been developed for cosmetic ingredients (Dent et al., 2018; Gilmour et al.,

121 2020). Methods for incorporation of mechanistic studies as supporting evidence in hazard

122 and/or risk assessment is included in the U.S. NTP OHAT handbook for systematic reviews,

123 the ORD staff handbook for developing IRIS assessments, and the draft TSCA interpretation

124 of systematic review methods to support chemical risk evaluations (EPA, 2022; EPA, 2023;

125 NTP OHAT, 2019). However, there is currently no complete framework for evidence-based

126 chemical risk assessment that integrates NAMs to facilitate the transition from use of

127 animals to the use of NAMs in chemical risk assessments.

"Next generation risk assessment in practice" is a project in the European Partnership for the Assessment of Risks from Chemicals (PARC). 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 task focusing on facilitating regulatory acceptance and use of NAMs. PARC is a seven-year partnership under Horizon Europe, including close to 200 institutions from 28 countries working in the areas of the environment or public health, and three EU authorities (PARC, 2023). With the "Next generation risk assessment in practice" project, we aim to contribute to the development of a framework for evidence-based use of data generated by in vitro studies in human health hazard identification and characterisation by creating tools and guidance's. A webpage giving an overview of the planned work in the "Next generation risk assessment in practice" project has been created (VKM, 2023). The first step in our PARC project is to develop a tool for evaluation of internal validity for in vitro studies. The next steps, all focusing on in vitro studies, will be development of a tool for evaluation of external validity, creation of a guidance for evaluation of certainty in the evidence, and creation of guidance's for the identification of point of departure and the uncertainty in the point of departure. We chose to start focusing on creation of tools for validity assessment, as validity assessment is one of the critical steps in the systematic review process. Further, we chose to start focusing on in vitro models as there is a general agreement that these are important as replacement for animal studies to provide information for hazard/risk assessment (ECHA, 2016; EFSA et al., 2022; EPA, 2018) in a wider integrating approach. It has been suggested that in vitro models could be more suitable than animal models for the prediction of toxicity. For example, in vitro data did predict liver toxicity caused by the drug troglitazone whereas neither published animal nor human studies were able to accurately predict the hazard (Dirven et al., 2021).

Several in vitro study designs exist; however, we have chosen only to focus on cell culture studies (meaning studies using cells derived from multicellular organisms). This delimitation is mainly due to feasibility, especially concerning the user testing, where the number of user testing participants will have to be very large to be able to test that the tool works on all types of in vitro study designs.

The implementation of this tool might be of help to improve the inclusion of NAMs in the chemical risk assessment process and facilitate regulatory uptake, with a focus on risk assessors' daily practice and workflow.

The first step is the development of a tool for evaluation of internal validity of in vitro studies. While many tools have been created for assessing in vitro studies, no single tool isthere is a

163 priori lack of consensus on developing a tool with the application of rigorous methods.

164 ~~obviously authoritative specifically for assessing the internal validity of *in vitro* study designs.~~

165 We therefore aim to address this situation by ~~developing a new tool for evaluating the~~

166 ~~internal validity of *in vitro* studies,~~ using methods that ensure we are building on prior work,

167 with a degree of rigour consistent with our intent to provide an authoritative assessment

168 tool. ~~Internal validity is the extent to which a study is free from bias, where bias is~~

169 ~~"systematic error, or deviation from the truth, in results" (Cochrane Collaboration, 2005). A~~

170 ~~test performed *in vitro* ("in the glass") means that it is done outside of a living organism and it~~

171 ~~usually involves isolated tissues, organs or cells (ECHA, 2023).~~ We also intend to use the

172 findings of INVITES-IN to prepare a guidance on the design and conduct of *in vitro* studies

173 that will help researchers minimise and/or transparently identify potential biases in their

174 studies.

1.2~~1~~ Objective

176 The aim of this project is to create INVITES-IN, a tool for evaluating the internal validity of *in*

177 *vitro* studies. ~~While t~~The INVITES-IN tool will be designed specifically to be applied to cell

178 culture models (e.g. cell lines, primary cell models, co-cultures, monolayer and 3-D cell

179 models systems) treated with a single-chemical substance exposure, measuring any

180 outcome~~s, we intend that the tool be applicable (potentially with modification) to other *in vitro*~~

181 ~~study designs or other NAMs such as organ-on-a-chip, *in ovo, fish embryos, ex vivo, in*~~

182 ~~*chemico,* etc., and chemical mixture studies.~~ We anticipate that the tool will be applicable

183 (potentially with modification) to other *in vitro* study designs or other NAMs such as organ-

184 on-a-chip, *in ovo, fish embryos, ex vivo, in chemico,* etc~~.~~, and chemical mixture studies, but

185 this will not be addressed in this study.

186 To contribute to its usability, INVITES-IN will be accompanied by instructions to guide the

187 user through the evaluation of internal validity of *in vitro* studies step-by-step.

188 While there is good empirical evidence from several domains that certain features of how a

189 study is designed, conducted, and analysed can introduce bias, it is usually not possible to

190 determine how much bias a given feature has introduced on any specific occasion (Savović

191 et al., 2012). INVITES-IN therefore follows conventional guidance (Boutron et al., 2022;

192 Frampton et al., 2022) in being designed to differentiate studies with relatively higher risk of

193 bias from studies with relatively lower risk of bias.

1.3~~2~~ Project governance

195 The development of INVITES-IN is part of the PARC project "Next generation risk

196 assessment in practice" [Project 101057014 – PARC]. A project group (PG) has been

197 established with the responsibility for developing and implementing the tool for evaluation of

198   internal validity of *in vitro* studies. The project is led by the Norwegian Institute of Public
199   Health represented by the Norwegian Scientific Committee for Food and Environment
200   (Norway). The project partners are Benaki Phytopathological Institute (Greece), Istituto
201   Superiore di Sanità (Italy) and the University of Basel (Switzerland).

202   A scientific advisory group (SAG) consisting of experts in systematic review principles,
203   chemical risk assessment, toxicology, NAMs, and/or methods for tool development, several
204   of whom have been directly involved in developing approaches to assessing the validity of *in*
205   *vitro* studies, has been established. The SAG gives strategic guidance and support to the
206   PG and share information about ongoing projects addressing similar questions to ensure
207   that the outcome of this project complements and builds on the work of others and thereby
208   creates synergies and avoids duplication of efforts.

209   ## 2 Materials and methods

210   ### 2.1 Study design

211   #### 2.1.1 An overview of the creation of INVITES-IN

212   The method for creating INVITES-IN will follow the general framework for developing quality
213   assessment tools suggested by Whiting et al. (2017). This is a broad framework of general
214   principles rather than a tightly-prescribed standard but gives the general structure of our
215   approach. Four studies will be performed to create INVITES-IN (Figure 1). This protocol
216   describes Studies 1, 2 and 3, and the timeline is shown in Figure 2. A separate protocol will
217   be prepared for Study 4.

218   The tool will consist of signalling questions and criteria for reaching risk-of-bias judgments
219   for each signalling question. Criteria are the issues that have to be fulfilled to avoid bias.
220   Signalling questions are questions that the users of the tool answer in order to determine
221   whether the criteria have been fulfilled. The technical solution for the tool has not yet been
222   decided; however, we intend to make an online tool.

223
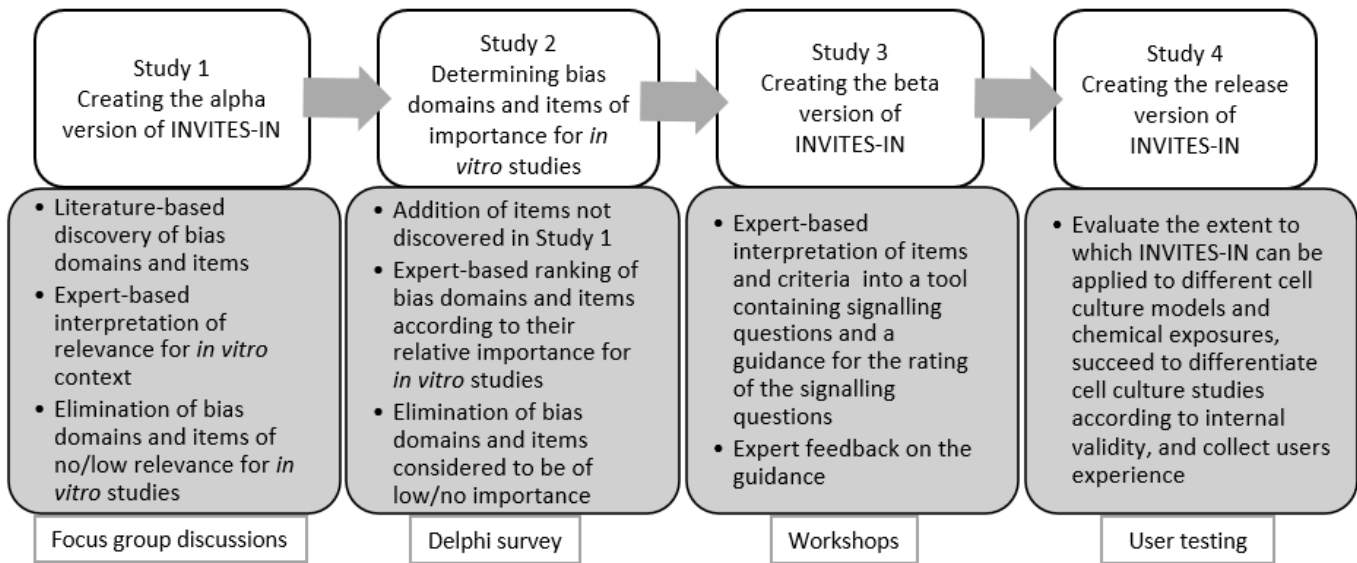
Figure 1. An overview of the four studies that will be performed to create the release version of INVITES- IN.

226

227

228

Figure 2. An overview of the 2023/24 timeline for the creation of the beta version of INVITES-IN.

All studies include participation of experts. We aim to achieve diversity among the participants by including scientists from different fields, professional backgrounds, geographical locations, gender, and having experience with different cell culture models. The

target group for the use of the tool (i.e., end-users) includes *in vitro* scientists and risk assessors conducting literature reviews in hazard assessments/safety evaluations, which could be part of a chemical risk/safety assessment, a systematic review or both, for regulatory or research purposes.

To get the input we need to develop the tool, we aim to recruit participants experienced with *in vitro* research that are representative for the end-users. For the studies 1 and 3, we aim to recruit some participants also having experience with systematic reviews, some also having experience with chemical risk assessment, and some having no experience with systematic reviews or chemical risk assessments. For study 2, we consider it critical that all participants have systematic review experience, as this is the study where the importance of different internal validity items will be ranked. Previous experience with evaluation of internal validity is considered important to be able to rank importance of different internal validity items. All groups of end-expected users are covered by the networks of the PG and the SAG. Potential participants will therefore be identified through nomination by PG and SAG members, who will be requested to nominate three potential participants. For each nominated participant, an overview of their scientific expertise and experience, affiliation, geographical location, and gender will be prepared. From the pool of nominated participants, PG will select participants that will be invited. In the selection process, PG will ensure diversity among the participants by including scientists from different fields having different professional backgrounds and experience with different cell culture models, covering a variety of geographical locations, and having an even gender distribution. In each focus group, all participants should be affiliated with different institutions, located in at least four different countries. This way we will avoid having an overrepresentation of focus group participants from a few institutions or from a too limited number of countries. We consider that this described process will make it possible to carry out the recruitment without it being an overly time-consuming process, and at the same time secure sufficient diversity in the group of participants.

The tasks and workload for the participants, the outcome of their contribution, and the participant eligibility criteria, are shown in Figure 3 and Table 1. Note that it is not expected that the same persons participate in all studies. It is planned that the persons participating in Study 1 will be also invited to participate in Study 3.

For all three studies, the potential participants will receive information about the project when they are contacted by email, and participants that accept the invitation will be requested to complete a declaration of interest form. The PG will evaluate the declaration of interest forms, focusing mainly on identification of potential conflicts of interest that may interfere with the participants contribution and role in the focus group discussion.

9

269 Previous studies report average or median time for the assessment of RoB of a study to
270 range from 20 to 40 minutes (Eick et al., 2020; Momen et al., 2022). We intend to keep the
271 time needed for assessment of one cell culture study within this range.

272 All data analyses will be done by the PG members. All raw data from each study will be
273 anonymised and made available as supplementary to the respective publications.

274

Study 1
Focus group discussions 18-24 participants (12 is the minimum).
•All participate in two online focus group discussions (~90 min for each discussion).
•Three focus groups in total (two is the minimum; six to eight participants per group).
•A PG member leads the discussion.
Outcome
• Items and criteria relevant for *in vitro* studies are identified.
• Items of no/low relevance for *in vitro* studies are eliminated.

Study 2
Delphi survey 20-30 participants (15 is the minimum).
•All complete two online surveys (~4 h).
•At least 10 participate in the online guided discussion (~60 min).
•A PG member leads the guided discussion.
Outcome
•Items and criteria important for introduction of bias to *in vitro* studies are identified.
•Items considered to be of low/no importance are eliminated.

Study 3
Workshops 5-24 participants.
•The participants from study 1 will be invited.
•All participate in one online workshop (~60 min).
•A PG member leads the workshop.
Outcome
•Strengths and weaknesses of the guidance document are identified.
•Feedback and suggestions for revisions are received.

275

276 **Figure 3.** Participants' tasks and workload in Study 1 to 3, and the outcome of their
277 contribution.

278

279 **Table 1**. An overview of the criteria for participation in Study 1 to 3.

Selection of participants		Study 1	Study 2	Study 3
Scientific experience and expertise	*In vitro* models	x		x
	In vitro models AND chemical risk assessment	x		x
	In vitro models AND systematic review methods	x	x	x
	In vitro models AND experienced with the development of relevant guidance documents for chemical risk assessors	x		x
Balancing factors	Academia	x	x	x
	Governmental institutions (including risk assessment institutions and research institutes)	x	x	x
	Private sector research institutions	x	x	x
	Gender distribution	x	x	x
	Demographic distribution	x	x	x
	Regional distribution	x	x	x
Academic level	Post-doctoral level or higher	x	x	x
Language	English, level B1 or higher	x	x	x

280

281 2.1.2~~1~~ Ethical review

282 Ethical approval has been given by the Norwegian Institute of Public Health.

283 2.2 Study 1: Creating the alpha version of the tool

284 2.2.1 Introduction and objective

285 The objective of Study 1 is to create a straw-man or alpha version of INVITES-IN that can be

286 further developed via a modified Delphi process (see section 2.3.2 for description). In Study

287 1, a list of characteristics of the design, conduct, and analysis of an *in vitro* study that can

288 introduce bias into its results or findings will be compiled, organised thematically, and then

289 interpreted into a draft set of structured signalling questions that constitute the alpha version

290 of INVITES-IN.

291 The knowledge goal is to have the expert interpretations of the relevance of bias domains

292 and items for *in vitro* studies.

A pilot focus group discussion was arranged to get an impression of the time needed for the focus group discussions, to test the technical functions, and to get feedback on factors related to the presentation of questions and the use of examples that may be of importance to conduct successful focus group discussions.

2.2.2 Method

We will include three focus groups with six to eight participants in each group (Figure 3).

An overview of the workflow and the responsibilities in Study 1 are given Table 2.

Table 2. An overview of Study 1.

Phase	Task	Responsible
Plan	Prepare the list of bias domains and items.	Project group
	Create questions for the focus group discussions.	
	Define inclusion criteria for focus group participants.	Project group and scientific advisory group
	Nominate and recruit focus group participants fulfilling the inclusion criteria.	
Actions	Carry out the focus group discussions.	Project group
	Analyse results and prepare the final report.	
Result	Bias domains and items of relevance for *in vitro* studies are identified and included in the alpha version of the tool.	Project group

2.2.2.1 Identifying relevant bias domains and items

A list of bias domains and items of potential relevance for *in vitro* studies will be prepared using several literature sources. This list will serve as a starting point for the creation of INVITES-IN and provide the basis for the focus group discussions. The literature sources are as follows: two systematic reviews on validity tools for *in vitro* models (Tran et al., 2021; Whaley et al., in preparation)(Whaley et al., in preparation), a publication on study sensitivity that includes assessment items that may relate to internal validity but may not be included in other tools (Cooper et al., 2016), and tools for evaluation of risk of bias (EPA, 2022; NTP OHAT, 2015; NTP OHAT, 2019; Roth et al., 2021; Sterne et al., 2019).

2.2.2.2 Focus group participants

Eligible focus group participants will be scientists with or without systematic review experience that are active in the field of *in vitro* research in academia, governmental institutions (including risk assessment institutions and research institutes) or private research institutes, at post-doctoral level or higher, and level B1 English speakers (see Table 1). PG

316 and SAG will nominate participants. We aim to have an equal gender distribution, a
317 reasonable demographic and regional distribution, and a group size of six to eight
318 participants as this group size is recommended to generate diverse ideas but not so many
319 participants that they do not have a chance to share perspectives (Krueger et al., 2001). The
320 minimum number of participants in a focus group is considered to be four. All participants in
321 a focus group will be affiliated with different institutions in an attempt to achieve variation in
322 input and perspective, and they should be working with a variety of *in vitro* models to cover a
323 wide range of experimental systems. No compensation is offered for the participation, and
324 participants will not be offered co-authorship.

325 Potential focus group participants will be contacted via email. They will receive a document
326 with information about the project, the purpose of the focus groups and the focus group
327 discussions, that the use of information learned in the meeting will not allow for identification
328 of the focus group participants, the withdrawal procedure, the financial source, and the
329 approximate time for the focus group meeting. Focus group participants must actively
330 confirm their consent by email.

331 We aim to have three different focus groups (Krueger et al., 2001), however, two groups are
332 considered to be the minimum. All groups will be presented with the same information and
333 questions, although the direction in which discussion is steered may depend on how
334 comprehensively previous focus groups were able to cover each issue. The need for
335 including an additional group will be discussed if new insights are presented during the
336 meetings, or if areas needing discussion were not addressed.

337 *2.2.2.3 Focus group discussion*
338 We plan to have two group discussions per focus group. The second meeting will be
339 cancelled if considered not to be needed. The discussions will be carried out as online
340 meetings and will be recorded. A PG member will act as a focus group moderator and lead
341 the discussions in the meeting, and another PG member will handle the logistics (the
342 assistant moderator).

343 The complete list of identified bias domains and items will be the starting point for the focus
344 group discussions. The discussions will be facilitated with a view to addressing twothree
345 questions (numbering is for referencing purposes and the questions will not necessarily be
346 presented in this order):

347 1. Are there any gaps in the identified domains or items that could influence systematic
348 error in an *in vitro* study?
349 2. What characteristics of the design, conduct, or analysis of an *in vitro* study could
350 introduce systematic error into its results or findings?

3. Should INVITES-IN ask users to judge risk of bias directly under a set of organising
352 domains, or ask users about the presence or absence of study characteristics from
353 which a bias judgement can be extrapolated?

354 Question (1) will be addressed both by asking directly and inferred from analysis of the
355 discussion (see section 2.2.2.4 below). Question (2) will be directly asked.

356 Question (3) is being asked because the traditional approach to risk of bias assessment has
357 been to structure signalling questions according to the relevant bias domains (Higgins et al.,
358 2011; NTP OHAT, 2019), whereas the signalling questions in some more recent tools are
359 based on study characteristics and structured around whether the bias is introduced before,
360 during or after the exposure of the experimental system to the test item (i.e. prior, during and
361 after the administration of the chemical substance in the experiment), from which a risk of
362 bias judgement is extrapolated (Sterne et al., 2019).

363 Discussion relating to questions (1) and (2) will be structured in terms of the bias domains
364 defined in the Scientific Evidence Code System (SEVCO) (Table 3) (Alper et al., 2021b). The
365 SEVCO domains are chosen because they are consistent with the bias domains of Whaley
366 et al. (in prep) and the OHAT tool (NTP OHAT, 2019) but represent a more recent
367 normalised list of bias categories derived from a robust grounding and consensus process
368 (Alper et al., 2021a). These definitions are developed for human studies, and the relevance
369 for *in vitro* studies will be discussed in the focus groups. We acknowledge that not all bias
370 domains presented in Table 3 may be of relevance for *in vitro* studies. However, we will
371 include all bias domains with approved SEVCO definitions in the focus group discussions in
372 order to collect expert feedback on the relevance for *in vitro* studies. SEVCO draft bias
373 domains that have not been approved are not listed. Participants may suggest additional
374 bias domains.

375 **Table 3**2. Bias domains with approved definitions in the Scientific Evidence Code System
376 (FEvIR Platform Version 0.80.0, 06.12.2022).

Bias Domain	Definition	SEVCO code reference
Selection Bias	A bias resulting from methods used to select subjects or data, factors that influence initial study participation, or differences between the study sample and the population of interest	SEVCO:00002
Confounding Covariate Bias	A situation in which the effect or association between an exposure or outcome is distorted by	SEVCO:00016

	another variable. For confounding covariate bias to occur the distorting variable must be (1) associated with the exposure and the outcome, (2) not in the causal pathway between exposure and outcome, and (3) unequally distributed between the groups being compared.	
Performance Bias	A bias resulting from differences between the received exposure and the intended exposure	SEVCO:00017
Attrition Bias	A bias due to absence of expected participation or data collection after selection for study inclusion.	SEVCO:00019
Detection Bias	A bias due to distortions in any process involved in the determination of the recorded values for a variable.	SEVCO:00020
Analysis Bias	A bias related to the analytic process applied to the data.	SEVCO:00021
Reporting Bias	A bias due to distortions in the selection of or representation of information in study results or research findings.	SEVCO:00023
Early Study Termination Bias	A bias due to the decision to end the study earlier than planned.	SEVCO:00370

377

378 Focus group participants will be shown and have read to them the definitions for each bias

379 domain. Participants will then be led in discussion of how the domain might be active in the

380 *in vitro* context, with examples from their practical research experience of how systematic

381 error can be introduced into an *in vitro* study. For each bias domain, one example for animal

382 studies and one example for *in vitro* studies will be prepared and these will be presented

383 when there is a need for further clarification to start the discussion. Each focus group

384 meeting will last about 90 minutes, with approximately 10 minutes given to each domain.

385 Participants will be given an option to send additional thoughts and considerations on the

386 relevance of the discussed bias domains and items for *in vitro* studies to the PG by email

387 within a week after the focus group discussion.

388 *2.2.2.4 Data analysis and reporting*

389 Focus group transcripts will be analysed for potential risk of bias criteria and items that could

390 be added to the alpha version of INVITES-IN. For time efficiency, transcripts of the focus

391 group discussions will be machine-generated. Errors in transcription will only be corrected

when they affect coding and interpretation of the discussion and will be done by the focus group moderator and the assistant moderator. Anonymised transcripts will be shared as raw data and be included as supplementary materials. The original recordings, as they contain personally identifiable information, will not be made available.

The focus group transcripts will be annotated (coded) in order to provide qualitative data on the following: preferences of the participants for traditional versus more recent approaches to structure risk of bias assessment ("preferred approach"), including reasons for and against; the participants' ideas about how researchers' approaches to designing, conducting, analysing and reporting studies ("issues") can potentially introduce systematic error, including their potential importance; the participants ideas about when ("time-point") systematic error is introduced; the participants ideas about the relevance ("relevance") for *in vitro* studies.

Data on preferred approach, issues, time-points, and relevance will be annotated by two investigators with a high level of expertise in bias assessment working independently then reconciling their coding decisions in discussion with a third investigator with experience in coding and reconciliation. The annotation environment will be ~~EPPI-Reviewer (version 4.13.0.2) (EPPI Centre, 2022; Thomas, 2022)~~Microsoft Word. The annotators will reach consensus for coding using the codebook through coding a part of one transcript together and discussing differences in interpretation~~will be trained on a page of transcript~~, and ~~where~~ they will agree on the rules for annotation (e.g., sentence or word highlighting for codes) and document these as their coding strategy in a coding manual.

Coding will be a mix of deductive (prespecified) and inductive (ad hoc) annotation. The definitions of the deductive codes are included in Table 4. ~~Code Book below (Table 4) gives the definitions of deductive codes and~~, and we have also indicate~~ds~~ where we already anticipate that codes will be developed inductively, though further inductive codes will be developed as needed. The Code Book is shown in Table 5. A report of the results of the annotation exercise, as a set of excerpted text strings aggregated under code categories and labelled with specific codes, will be generated as data for supporting development of the alpha version of INVITES-IN.

Table 4. The definition of the codes in the Code Book.

Code Category	Code	Definition
Issue	- Selection	An issue relating to selection bias
	- Confounding	An issue relating to confounding covariates bias

	- Performance	An issue relating to performance bias
	- Attrition	An issue relating to attrition bias
	- Detection	An issue relating to detection bias
	- Analysis	An issue relating to analysis bias
	- Reporting	An issue relating to reporting bias
	- Early Termination	An issue relating to early termination bias
	- [ad hoc codes]	Ad hoc codes will be created to classify limitations that do not fit into any of the prespecified bias categories (inductive coding)
Time-point	- Before exposure	An issue that may affect potential for systematic error prior to the exposure (administration of the chemical substance) in the experiment
	- During exposure	An issue that may affect potential for systematic error during the exposure (administration of the chemical substance) in the experiment
	- After exposure	An issue that may affect potential for systematic error after the exposure (administration of the chemical substance) in the experiment
Relevance	- Higher relevance	Argument or observation that an issue that may affect potential for systematic error is of potentially higher relevance
	- Lower relevance	Argument or observation that an issue that may affect potential for systematic error is of potentially lower relevance

422

423 **Table 5.** The Code Book.

Level 1	Level 2
Selection	- Before exposure
	- During exposure
	- After exposure
	- Higher relevance
	- Lower relevance
- Confounding	- Before exposure
	- During exposure

	- After exposure
	- Higher relevance
	- Lower relevance
- Performance	- Before exposure
	- During exposure
	- After exposure
	- Higher relevance
	- Lower relevance
- Attrition	- Before exposure
	- During exposure
	- After exposure
	- Higher relevance
	- Lower relevance
- Detection	- Before exposure
	- During exposure
	- After exposure
	- Higher relevance
	- Lower relevance
- Analysis	- Before exposure
	- During exposure
	- After exposure
	- Higher relevance
	- Lower relevance
- Reporting	- Before exposure
	- During exposure
	- After exposure
	- Higher relevance
	- Lower relevance
- Early Termination	- Before exposure
	- During exposure
	- After exposure
	- Higher relevance
	- Lower relevance
- [ad hoc codes]	- Before exposure
	- During exposure

		- After exposure
		- Higher relevance
		- Lower relevance

2.2.3 Results and outcome

The focus group participants will not make decisions but provide ideas and recommendations. Their feedback on ~~approach preference,~~ issues, time-points, and relevance for the *in vitro* context will be used by the PG to prepare the alpha version of INVITES-IN, which will contain all bias domains and items considered to be of relevance for *in vitro* studies with reasonings. The final decisions regarding the inclusion of bias domains and items in the alpha version of INVITES-IN will be made by the PG members involved in this study. An overview of bias domains and items that are not included in the alpha version will be included in the study report, and comprehensively documented in supporting data. The intent here is not to permanently exclude any items, but to generate a list of practical length for analysis by the modified Delphi process. Decisions about exclusion of domains or items at this stage affect only the alpha version of INVITES-IN and are not final: if the Delphi process reintroduces any excluded concepts, this will supersede the initial decision made by the PG.

2.3 Study 2: Determining bias domains and items of importance for *in vitro* studies

2.3.1 Introduction and objective

The objective is to eliminate, add to, or refine the proposed bias domains and assessment items that are generated by Study 1. This provides the final data to be interpreted into the beta version of INVITES-IN in Study 3.

The feature tested is the importance of the bias domains and items included in the alpha version of INVITES-IN for the internal validity of *in vitro* studies.

The knowledge goal is to have the expert interpretations of the importance of bias domains and items for *in vitro* studies.

2.3. 2 Method

A modification of the Delphi technique (Dalkey and Helmer, 1963) will be used to obtain subjective opinions on the importance of bias domains and items for *in vitro* studies from experts experienced with both *in vitro* studies and systematic review principles. The Delphi technique gives the opportunity to collect subjective expert statements anonymously and gives the desired transparency, without e.g. social or personality-based factors resulting in

454 one expert's feedback influencing the feedback another expert in the group. Therefore, this
455 approach is considered to be an appropriate technique to identify expert agreement.

456 A two-round digital Delphi survey will be conducted, followed by an online workshop for
457 guided discussions. In both rounds, expert panellists will complete a questionnaire. From
458 each Delphi round, the outcome will be subjective expert feedback on importance of bias
459 domains and items, and we will use these data to identify expert agreement on bias domains
460 and items important for internal validity of *in vitro* studies. Bias domains and items for which
461 agreement were not reached during the two Delphi rounds will be discussed in the
462 workshop. In addition, the participants will be asked to give input on the wording of the
463 questions in each Delphi round and during the guided discussion.

464 An overview of the workflow and the responsibilities in Study 2 is given Table 65.

465 **Table 65.** An overview of Study 2.

Phase	Task	Responsible
Plan	Define inclusion criteria for Delphi participants (expert panellists).	Project group
	Nominate and recruit expert panellists fulfilling the inclusion criteria.	Project group and scientific advisory group
	Create the questionnaire addressing the bias domains and items relevant for *in vitro* studies identified in Study 1.	
Actions	Delphi round 1 Expert panellists complete the questionnaire and have the possibility to suggest additional bias domains and items.	Project group
	Between Delphi round 1 and 2 Analyse results from round 1. Feedback from round 1 is given to the expert panellists. Bias domains and items which met criteria for identification of agreement for inclusion in INVITES-IN are removed. Bias domains and items which met criteria for identification of agreement for exclusion from INVITES-IN are removed.	

	New questions may be included, existing questions may be revised.	
	Delphi round 2 Feedback from round 1 is given to the expert panellists. Expert panellists complete the questionnaire. Analyse results from round 2.	
	Workshop Expert panellists will be guided through a discussion of uncertainties related to bias domains and items for which agreement for inclusion or exclusion were not identified. Prepare transcripts, organise, and summarise results.	Project group
Result	Expert agreement on bias domains and items of importance for internal validity of *in vitro* studies is identified.	Project group

466

2.3.2.1 Delphi participants

468 Eligible Delphi participants will be scientists that are active in the field of *in vitro* research and

469 have some experience with systematic literature review principles, are affiliated in academia,

470 governmental institutions (including risk assessment institutions and research institutes) or

471 private research institutes, at post-doctoral level or higher, and level B1 English speakers

472 (see Table 1). PG and SAG will nominate participants.

473 We aim to have an even gender and geographical location distribution for the potential

474 participants that are invited to participate. The number of participants will be 20 to 30 (see

475 Figure 3), depending on the number of suitable candidates identified by PG and SAG and

476 the candidate's willingness to participate. The minimum number of participants is considered

477 to be 15.

478 Potential participants will be contacted via email, and they will receive a letter with

479 information about the project and the purpose of the Delphi survey including the fact that the

480 use of individual survey responses will not allow for identification of the participant, the

481 withdrawal procedure, the financial source, as well as the approximate time for completion of

482 the questionnaires. Participants must actively confirm their consent by email to be included

483 as a participant. Before each Delphi round and the guided discussion, participants will

484 receive instructions. Participants are eligible to be co-authors of the Delphi study manuscript

485 if they also read and comment on the final draft. No compensation or other incentives are

486 offered for the participation.

487   *2.3.2.2 Delphi rounds and workshop with guided discussion*

488   A Delphi round is defined as the process where the expert panellists complete a

489   questionnaire. Before each round, expert panellists will receive a document with information

490   about the project, the Delphi survey, and how the Delphi questionnaire information will be

491   handled and used.

492   The PG develops the questionnaire based on the alpha version of INVITES-IN prepared in

493   Study 1. The questionnaire will be prepared as an Excel form, and it will be sent to the

494   expert panellists by email. The expert panellists rate the importance of different bias

495   domains and items for the internal validity of *in vitro* studies. A 5-point Likert scale, with the

496   categories "strongly disagree" (1), moderately disagree (2), neutral (3), moderately agree (4),

497   and strongly agree (5) is used as response options (Verhagen et al., 1998).

498   The expert panellists will have two weeks to complete the questionnaire in each Delphi

499   round, and they will receive up to three email reminders to complete each round. Panellists

500   not responding within the deadline in one of the two Delphi rounds will be excluded from that

501   round. Removed participants will not be replaced. Participants excluded from the first round

502   will also be excluded from the second round.

503   **Delphi round 1**: The questionnaire is completed by the expert panellists, and they will also

504   be able to suggest additional bias domains and items and alternative wording.

505   **Between Delphi round one and two**:

506   -   The results are analysed, and expert panellists receive feedback on average rating
507   and distribution of ratings of importance of bias domain and items.
508   -   The questionnaire is revised. Bias domains and items which met criteria for
509   identification of agreement for inclusion or exclusion from INVITES-IN are removed.
510   New questions may be included, existing questions may be revised.

511   **Delphi round 2**: The revised questionnaire is completed by the expert panellists.

512   **Between Delphi round two and the workshop**:

513   -   Results are analysed, and expert panellists receive feedback on average rating and
514   distribution of ratings of importance of bias domain and items.
515   -   Bias domains and items that did not reach agreement for either inclusion or exclusion
516   in round two are included in the guided discussion workshop.
517   -   An overview of all bias domains and items that did not reach agreement for either
518   inclusion or exclusion will be prepared and sent to the expert panellists who will be
519   requested to include arguments for considering the items to be of higher or lower

520   importance. PG will prepare an overview of all arguments, which will be sent to
521   workshop participants.

522   *Workshop: A workshop will be arranged to have a guided discussion on items where no*
523   *agreement on importance for in vitro studies has been identified. The starting point for the*
524   *discussion of each of these items will be the overview of arguments created between the*
525   *Delphi round two and the workshop. During the discussion, we will ask the participants to*
526   *give reasonings for agreeing or disagreeing with the arguments. New arguments that*
527   *emerge from the guided discussion will be included in the overview. A PG member will lead*
528   *and moderate the guided discussion. The workshop will be recorded and transcripts from the*
529   *workshop will form the basis for the revision of the list of arguments.*

530

531   *2.3.2.3 Data analysis and reporting*

532   One PG member will send out the questionnaires, receive the completed questionnaires
533   from the expert panellists, and anonymise the answers. This person will not be involved in
534   the data analysis.

535   Expert panellist characteristics such as gender distribution and geographic localisation will
536   be reported. The response rate (percentage) for expert panellist completing the Delphi
537   survey will be calculated and reported. The average group response, changes in rating
538   between rounds, as well as modifications of the questionnaire, will be reported. The expert
539   panellists rating of the questions will be analysed independently for round one, round two,
540   and the guided discussion, and median, mean, standard deviation and the interquartile
541   range will be reported.

542   Criteria for identification of agreement in round one and two:

543   - Agreement for inclusion of bias domains and items is identified when 70% of the
544   expert panellists rate the relevance and wording of a question as the category
545   "moderately agree" or "strongly agree" (1 and 2 on the 5-point Likert scale).
546   - Agreement for exclusion of bias domains and items is identified when 70% of the
547   expert panellists rate the relevance of a question as the category "moderately
548   disagree" or "strongly disagree" (1 and 2 on the 5-point Likert scale).

549   Decisions on identification of agreement will be made by the PG members involved in this
550   study.

551   The transcripts from the workshop will be anonymised and made available as supplementary
552   materials.

2.3.3 Results and outcome

Study 2 will result in a list of bias domains and items i) for which there were agreement that the domain or item is of importance when evaluating risk of bias of *in vitro* studies, ii) for which there were agreement that the domain or item is not of importance when evaluating risk of bias of *in vitro* studies, and iii) where agreement was not reached for either inclusion or exclusion in the two rounds of Delphi or in the guided discussion. For the items where agreement was not reached, arguments for considering a given item as higher or lower importance will be included.

2.4 Study 3: Creating the beta version of INVITES-IN

2.4.1 Introduction and objective

The objective is to create the beta version of INVITES-IN, that will be advanced to user testing. This will consist of two elements: the tool itself, consisting of a set of signalling questions and a process for deriving a risk of bias assessment; and a guidance document explaining how to use the tool. The guidance document will also include relevant examples of ratings of cell culture studies. This will be given as short texts illustrating possible reporting in a publication together with explanations and reasonings for how this is intended to be rated when applying INVITES-IN.

The knowledge goal is to have a complete set of signalling questions addressing bias domains and items of importance for introduction of bias to *in vitro* studies and the criteria for the rating of the questions.

2.4.2 Method

An overview of the workflow and the responsibilities in Study 3 is shown in Table 7.

Table 7. An overview of Study 3.

Phase	Task	Responsible
Plan	Signalling questions are formulated. Guidance for rating the signalling questions is prepared. The process for compiling the results from the rating of the signalling questions into an overall assessment of the risk of bias for each study is created.	Project group
	Invite members of the focus group that interpreted bias domains and items for *in vitro* context (Study 1) to participate in an online workshop.	
Actions	Workshop	Project group

	Get feedback on the presentation of and information in the guidance document (Study 1).	
Result	The guidance is revised according to the workshop feedback. The beta version of INVITES-IN is finalised.	Project group

576

2.4.2.1 Draft version of INVITES-IN

578 The draft version of INVITES-IN will be prepared by the PG. The outcome of Study 2 will be
579 used to formulate the signalling questions. The guidance document will contain explanations
580 of how each signalling question should be rated.

2.4.2.2 Workshop participants

582 Members of the focus group participating in Study 1 will be invited to participate in an online
583 workshop, except for those who also participated in the Delphi process which will be
584 excluded. No compensation is offered for participation, and participants will not be offered
585 co-authorship.

2.4.2.3. Workshops

587 One or more online workshops will be arranged to collect feedback on both the presentation
588 and the information in the guidance document. Regarding the feedback on information in the
589 guidance document, the focus will be on the suggested criteria for the rating of the signalling
590 questions and whether we have succeeded in formulating these so that it is the factors that
591 are of considered to be of greatest importance for the introduction of bias that are given the
592 most weight.

593 We also attempt to collect feedback from the participants regarding the presentation of the
594 signalling questions from the workshops; whether they should be structured according to the
595 relevant bias domains or be based on study characteristics and structured around whether
596 the bias is introduced before, during or after the exposure of the experimental system to the
597 test item (i.e. prior, during and after the administration of the chemical substance in the
598 experiment).

599 When possible, the number of participants in a workshop will be six to eight. However,
600 workshops with fewer participants will be considered in order to facilitate participant
601 recruitment.

602 The workshops will be recorded.

2.4.2.4 Data analysis and reporting

Transcripts An overview of the feedback on the guidance document received in the workshops will be prepared and made available as supplementary materials. and used by the PG for the revision of INVITES-IN.

Based on the feedback from participants in the workshops, PG will make the final decision on the need for revision.

2.4.3 Results and outcome

The beta version of the tool is ready for user testing.

3 Discussion

This protocol describes the methodological approach for the development of the INVITES-IN tool. Few *in vitro* assessment tools have been developed using rigorous methods that include a literature review, Delphi process, and formal user testing; we are not aware of a single published tool for the assessment of the internal validity of *in vitro* studies that follows all the steps we have outlined in our protocol. In this protocol, we have proposed an approach similar to that of ROB2 (Sterne et al., 2019) and ROBINS (Sterne et al., 2016). The approach chosen fulfils the framework for developing quality assessment tools (Whiting et al., 2017), which is to our knowledge the only existing framework for how to develop quality appraisal tools. Although, we cannot be certain that the chosen approach is the best approach, we feel confident that the methods chosen are rigorous and have been agreed upon of more than 20 experienced experts/scientists. Also, we have focused on transparency and there detailed method descriptions and collected data (transcripts and more) will be made publicly available. is rigorously grounded and the tool development process fulfils the framework for developing quality assessment tools (Whiting et al., 2017). Our methodological approach comprises four separate studies and involves both focus groups, two-round Delphi survey and user-testing at different stages. A separate protocol will be prepared for the user testing (Study 4). Involving groups of experts in every study reduces the level of expert judgements made by the project group, and also ensure that the tool development is based on a wide range of feedback from experts that are the intended user of the tool. It might be that including more participants in the three studies described in this protocol would give additional interpretations of the relevance and importance of bias domains and items for *in vitro* studies.

It may be a challenge to recruit enough experts to ensure sufficiently powering of the studies. To facilitate the recruitment process, the workload for the participants is limited to the absolute minimum. Also, participants in the Delphi-survey, which are likely to have the

637 largest workload for the participants, will be offered authorship on the Delphi study
638 manuscript.

639 The described approach will not include the assessment of magnitude or direction of the
640 bias. We believe that these issues need to be addressed by empirical research in addition to
641 expert knowledge elicitation. We acknowledge the importance of assessing magnitude and
642 the direction of bias, however, the amount of work and time it will take to properly address
643 this, will not be possible at this stage of the tool development.

644 Given that assessment of *in vitro* studies is likely to become a fast-moving field, we
645 acknowledge there may be a need for the tool to be updated to reflect rapid changes in
646 consensus on how to do this, and/or it may be fast movement toward modifying INVITES-IN
647 for other specific NAM study designs. A plan for the update or modification of INVITES-IN is
648 not included in this protocol, as it is restricted to describe the process for the creation of this
649 tool.

650 ## Dissemination
651 A focus group interview report will be prepared.

652 A Delphi process report will be prepared, including the questionnaires used in round one and
653 round two.

654 The beta version of the tool, ready for user testing, will be prepared.

655 ## Abbreviations
656 NAM: new approach methodologies

657 PG: project group

658 SAG: scientific advisory group

659 ## Definitions
660 **Bias** are systematic errors, or deviations from the truth, in results or inference (Cochrane
661 Collaboration, 2005). For *in vitro* studies, systematic errors may be introduced in the study
662 design, conduction, and/or analysis, and cause the result to be an overestimate or
663 underestimate.

664 **Bias domains** are themes such as study performance, analysis, and reporting, under which
665 bias items can be organised/grouped.

666 **Bias items** are study properties that may be relevant for introduction of bias in results and/or
667 their interpretation. Criteria are the issues that have to be fulfilled for bias to be avoided. In

668   the guidance document for the INVITES-IN tool there will be criteria for reaching risk-of-bias
669   judgements for each signalling question.

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

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

674   **NAMs** have not yet a standard definition. However, there seems to be a general agreement
675   that the term "NAMs" include *in chemico*, *in silico* and *in vitro* studies. One established
676   definition is that NAMS includes any technology, methodology, approach, or combination
677   that can provide information on chemical hazard and risk assessment without the use of
678   animals, including *in silico, in chemico, in vitro*, and *ex vivo* approaches (ECHA, 2016; EPA,
679   2018).

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

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

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

694   Ethical considerations
695   Ethical approval has been given by the Norwegian Institute of Public Health.

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

References

Alper B.S., Dehnbostel J., Afzal M., Subbian V., Soares A., Kunnamo I., Shahin K., McClure R.C. (2021a) Making science computable: Developing code systems for statistics, study design, and risk of bias. Journal of Biomedical Informatics 115:103685. DOI: 10.1016/j.jbi.2021.103685.

Alper B.S., Dehnbostel J., Lehmann H., Whaley P., Wilkins K.J., Tufte J., Yurk R.A., Ojha N., Afzal M. (2021b) For the COVID-19 Knowledge Accelerator (COKA) Initiative. Scientific Evidence Code System Development Protocol. Created November 16, 2021. Last revised December 8, 2021. Available at: https://docs.google.com/document/d/1pzGLdyVCKcu3s2gfSfPpXDQLlQsFnLZR14ld w0nD1g0.

Boutron I., Page M.J., Higgins J.P.T., Altman D.G., Lundh A., Hróbjartsson A. (2022) Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3

733   (updated February 2022). Cochrane, 2022. Available from
734   www.training.cochrane.org/handbook.

735 Cochrane Collaboration. (2005) Glossary of Terms in The Cochrane Collaboration, Version
736   4.2.5 http://aaz.hr/resources/pages/57/7.%20Cochrane%20glossary.pdf.

737 Cooper G.S., Lunn R.M., Ågerstrand M., Glenn B.S., Kraft A.D., Luke A.M., Ratcliffe J.M.
738   (2016) Study sensitivity: Evaluating the ability to detect effects in systematic reviews
739   of chemical exposures. Environ Int 92-93:605-10. DOI: 10.1016/j.envint.2016.03.017.

740 Dalkey N., Helmer O. (1963) An Experimental Application of the DELPHI Method to the Use
741   of Experts. Management Science 9:458-467. DOI: 10.1287/mnsc.9.3.458.

742 Dent M., Amaral R.T., Da Silva P., Amores, Ansell J., Boisleve F., Hatao M., Hirose A.,
743   Kasai Y., Kern P., Kreiling R., Milstein S., Montemayor B., Oliveira J., Richarz A.,
744   Taalman R., Vaillancourt E., Verma R., Posada N.V.O.R.C., Weiss C., Kojima H.
745   (2018) Principles underpinning the use of new methodologies in the risk assessment
746   of cosmetic ingredients. Computational Toxicology 7:20-26. DOI:
747   10.1016/j.comtox.2018.06.001.

748 Dirven H., Vist G.E., Bandhakavi S., Mehta J., Fitch S.E., Pound P., Ram R., Kincaid B.,
749   Leenaars C.H.C., Chen M., Wright R.A., Tsaioun K. (2021) Performance of
750   preclinical models in predicting drug-induced liver injury in humans: a systematic
751   review. Scientific Reports 11:6403. DOI: 10.1038/s41598-021-85708-2.

752 ECHA. (2016) New approach methodologies in regulatory science : proceedings of a
753   scientific workshop : Helsinki, 19-20 April 2016, European Chemicals Agency,
754   https://echa.europa.eu/documents/10162/21838212/scientific_ws_proceedings_en.p
755   df/a2087434-0407-4705-9057-95d9c2c2cc57.

756 ECHA. (2023) In vitro methods, European Chemicals Agency,
757   https://echa.europa.eu/support/registration/how-to-avoid-unnecessary-testing-on-
758   animals/in-vitro-methods.

759 EFSA, Tarazona J., Kass G., Dorne J.-L., Liem D., Paraskevopoulos K., Kleiner J., Heppner
760   C., Hugas M. (2022) Theme (Concept) Paper - New Approach Methodologies. EFSA
761   Supporting Publications 19:E200502E. DOI: 10.2903/sp.efsa.2022.e200502.

762 Eick S.M., Goin D.E., Chartres N., Lam J., Woodruff T.J. (2020) Assessing risk of bias in
763   human environmental epidemiology studies using three tools: different conclusions
764   from different tools. Syst Rev 9:249. DOI: 10.1186/s13643-020-01490-8.

765 EPA. (2018) Strategic plan to promote the development and implementation of alternative
766   test methods within the TSCA program. U.S. Environmental protection agency. EPA-
767   740-R1-8004, United States Environmental Protection Agency,
768   https://www.epa.gov/sites/default/files/2018-06/documents/epa_alt_strat_plan_6-20-
769   18_clean_final.pdf.

770 EPA. (2022) ORD Staff Handbook for Developing IRIS Assessments. U.S. EPA Office of

771 Research and Development, Washington, DC, EPA/600/R-22/268,

772 https://ordspub.epa.gov/ords/eims/eimscomm.getfile?p_download_id=545991.

773 EPA. (2023) Draft Protocol for Systematic Review in TSCA Risk Evaluations (assessed

774 10.02.2023), United States Environmental Protection Agency,

775 https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/draft-protocol-

776 systematic-review-tsca-risk-evaluations.

777 FEvIR Platform Version 0.80.0. (06.12.2022) Scientific Evidence Code System (SEVCO).

778 Computable Publishing®: CodeSystem Viewer,

779 https://fevir.net/resources/CodeSystem/27270#SEVCO:00001.

780 Frampton G., Whaley P., Bennett M., Bilotta G., Dorne J.-L.C.M., Eales J., James K., Kohl

781 C., Land M., Livoreil B., Makowski D., Muchiri E., Petrokofsky G., Randall N.,

782 Schofield K. (2022) Principles and framework for assessing the risk of bias for

783 studies included in comparative quantitative environmental systematic reviews.

784 Environmental Evidence 11:12. DOI: 10.1186/s13750-022-00264-0.

785 Gilmour N., Kern P.S., Alépée N., Boislève F., Bury D., Clouet E., Hirota M., Hoffmann S.,

786 Kühnl J., Lalko J.F., Mewes K., Miyazawa M., Nishida H., Osmani A., Petersohn D.,

787 Sekine S., van Vliet E., Klaric M. (2020) Development of a next generation risk

788 assessment framework for the evaluation of skin sensitisation of cosmetic

789 ingredients. Regulatory Toxicology and Pharmacology 116:104721. DOI:

790 10.1016/j.yrtph.2020.104721.

791 Krueger R.A., Casey M.A., Donner J., Kirsch S., Maack J.N. (2001) Designing and

792 Conducting Focus Group Interviews, Social Development Papers 31. Social Analysis:

793 Selected Toolds and Techniques, Washington, D.C: World Bank,

794 http://web.worldbank.org/archive/website01028/WEB/IMAGES/SDP_36.PDF.

795 Momen N.C., Streicher K.N., da Silva D.T.C., Descatha A., Frings-Dresen M.H.W., Gagliardi

796 D., Godderis L., Loney T., Mandrioli D., Modenese A., Morgan R.L., Pachito D.,

797 Scheepers P.T.J., Sgargi D., Paulo M.S., Schlünssen V., Sembajwe G., Sørensen K.,

798 Teixeira L.R., Tenkate T., Pega F. (2022) Assessor burden, inter-rater agreement

799 and user experience of the RoB-SPEO tool for assessing risk of bias in studies

800 estimating prevalence of exposure to occupational risk factors: An analysis from the

801 WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury.

802 Environment International 158:107005. DOI:

803 https://doi.org/10.1016/j.envint.2021.107005.

804 NTP OHAT. (2015) Protocol to evaluate the evidence for an association between

805 perfluorooctanic acid (PFOA) and perfluorooctane sulfonate (PFOS) exposure and

806 immunotoxicity, Office of Health Assessment and Translation (OHAT), Division of the

807 National Toxicology Program, National Institute of Environmental Health Sciences,
808 https://ntp.niehs.nih.gov/ntp/ohat/pfoa_pfos/protocol_201506_508.pdf. pp. 65-70.
809 NTP OHAT. (2019) Handbook for Conducting a Literature Based Health Assessment Using
810 OHAT Approach for Systematic Review and Evidence Integration, Office of Health
811 Assessment and Translation (OHAT), Division of the National Toxicology Program,
812 National Institute of Environmental Health Sciences,
813 https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf.
814 PARC. (2023) PARC_What we do_Thematic areas_News_Events_Scientific publications
815 (assessed 11.05.2023), https://www.eu-parc.eu/#news-and-events.
816 Roth N., Zilliacus J., Beronius A. (2021) Development of the SciRAP Approach for
817 Evaluating the Reliability and Relevance of in vitro Toxicity Data. Frontiers in
818 Toxicology 3:746430. DOI: 10.3389/ftox.2021.746430.
819 Savović J., Jones H.E., Altman D.G., Harris R.J., Jüni P., Pildal J., Als-Nielsen B., Balk E.M.,
820 Gluud C., Gluud L.L., Ioannidis J.P., Schulz K.F., Beynon R., Welton N.J., Wood L.,
821 Moher D., Deeks J.J., Sterne J.A. (2012) Influence of reported study design
822 characteristics on intervention effect estimates from randomized, controlled trials.
823 Ann Intern Med 157:429-38. DOI: 10.7326/0003-4819-157-6-201209180-00537.
824 Sterne J.A.C., Hernán M.A., Reeves B.C., Savović J., Berkman N.D., Viswanathan M.,
825 Henry D., Altman D.G., Ansari M.T., Boutron I., Carpenter J.R., Chan A.-W., Churchill
826 R., Deeks J.J., Hróbjartsson A., Kirkham J., Jüni P., Loke Y.K., Pigott T.D., Ramsay
827 C.R., Regidor D., Rothstein H.R., Sandhu L., Santaguida P.L., Schünemann H.J.,
828 Shea B., Shrier I., Tugwell P., Turner L., Valentine J.C., Waddington H., Waters E.,
829 Wells G.A., Whiting P.F., Higgins J.P.T. (2016) ROBINS-I: a tool for assessing risk of
830 bias in non-randomised studies of interventions. BMJ 355:i4919. DOI:
831 10.1136/bmj.i4919.
832 Sterne J.A.C., Page M.J., Savović J., Elbers R.G., Blencowe N.S., Boutron I., Cates C.J.,
833 Cheng H.-Y., Corbett M.S., Eldridge S.M., Hernán M.A., Hopewell S., Hróbjartsson
834 A., Junqueira D.R., Jüni P., Kirkham J.J., Lasserson T., Li T., McAleenan A., Reeves
835 B.C., Shepperd S., Shrier I., Stewart L.A., Tilling K., White I.R., Whiting P.F., Higgins
836 J.P.T. (2019) RoB 2: a revised tool for assessing risk of bias in randomised trials.
837 British Medical Journal 366:l4898. DOI: 10.1136/bmj.l4898
838 Tran L., Tam D.N.H., Elshafay A., Dang T., Hirayama K., Huy N.T. (2021) Quality
839 assessment tools used in systematic reviews of in vitro studies: A systematic review.
840 BMC Medical Research Methodology 21:101. DOI: 10.1186/s12874-021-01295-w.
841 Verhagen A.P., de Vet H.C., de Bie R.A., Kessels A.G., Boers M., Bouter L.M., Knipschild
842 P.G. (1998) The Delphi list: a criteria list for quality assessment of randomized

843   clinical trials for conducting systematic reviews developed by Delphi consensus. J

844   Clin Epidemiol 51:1235-41. DOI: 10.1016/s0895-4356(98)00131-0.

845 VKM. (2023) The Norwegian Scientific Committee of Food and Environment (VKM)

846   participates in the European Partnership for the Assessment of Risks from Chemicals

847   (PARC) (assessed 11.05.2023),

848   https://vkm.no/english/parc/parceuropeanpartnershipfortheassessmentofrisksfromche

849   micals.4.7205492a1864a8c8da2dcfd9.html.

850 Whaley P., Hooijmans C.R., Wattam S. (in preparation) Literature-based discovery of

851   assessment criteria for in vitro studies: a method and item bank.

852 Whiting P.F., Wolff R., Mallett S., Simera I., Savović J. (2017) A proposed framework for

853   developing quality assessment tools. Systematic Reviews 6:204. DOI:

854   10.1186/s13643-017-0604-6.

855