Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 109 (2022) 472–477

www.elsevier.com/locate/procedia

32nd CIRP Design Conference

# An open science platform for benchmarking engineering design researches

Romain Pinquié[a,*], Julien Le Duigou[b], Lou Grimal[c], Lionel Roucoules[d]

[a] Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France
[b] Mechanical Engineering department, Roberval Laboratory, Université de technologie de Compiègne, CS 60319 60200 Compiègne, France
[c] CD, CREIDD, Université de Technologie de Troyes, 12 rue Marie Curie, 10004 Troyes, France
[d] Arts et Metiers Institute of Technology, LISPEN, HESAM Université, F-13617 Aix-en-Provence, France

* Corresponding author. Tel.: +33-065-833-6305. E-mail address: romain.pinquie@grenoble-inp.fr

**Abstract**

In the last fifteen years, several studies showed that a large part of research in engineering design fails to provide objective evidence demonstrating real progress because of a lack of a generally accepted and effective validation environment to test, confirm or refute results. A better validation protocol would highlight the scientific added value on socio-technical systems, facilitate an incremental construction of solutions and improve the recognition of the engineering design research community. Thus, an open science environment for benchmarking engineering design research is proposed. First, a set of guidelines that helps researchers to define the minimum viable content of a benchmark is introduced. Second, the implementation of these guidelines is described in a web-based version-control and collaborative platform within which a community of researchers can engage in the building of a sustainable validation ecosystem. Future works will concentrate on the consolidation of the platform (benchmark examples, instructions, workflows, licenses, etc.) based on the feedback of the current members of the French S.mart special interest group (s-mart.fr) before advertising to a larger community of researchers.



*Keywords:* engineering design research; validation; open science; benchmark

## 1. Introduction

### 1.1. Context

In the last fifteen years, an increasing number of papers in engineering design research alert the community to a lack of focus on validation. The Design Society and, more recently, the CIRP STC-Dn community is also highlighting the lack of validation that make papers evaluated as low quality compared to other CIRP STCs [1].

**[Qualitative and quantitative]** First, a review of 80 % of ICED proceedings shows that research in engineering design heavily relies on engineering science and social science, but most research is a combination of empirical research, experimental research, development of new tools, implementation studies, and research dedicated to the theory and education [2]. The blend of qualitative and quantitative research approaches makes the validation difficult, as it requires mastering and implementing a mix of validation techniques.

**[Industry partners]** Second, research in engineering design is often motivated to answer specific problems that industry partners are facing. Thus, the validation of research requires evidence that supports the claim of improvements on problems that truly match challenges in the industry, which go beyond toy problems [3]. In addition, industrial partnerships force researchers to simultaneously work out solutions that are not only helpful to industrial partners but also contribute to the development of new generalizable design knowledge [4]. This twofold objective requires a tremendous amount of work to demonstrate "*what works, for whom, under which circumstances?*" [5], especially for the validation of the


10.1016/j.procir.2022.05.280

contribution to the industry that is often challenged by industrial situations that are too wide and too complex to provide objective evidence that justifies the claim [4] before the research project runs out of time and money [6,7].

**[Design methods]** Research studies that aim at improving design are mainly about the development of new design methods and tools [8]. The validation of both method and tools needs to be conducted separately [4]. However, the validation of methods is an embarrassing question for researchers because they "*need to be learned, interpreted and adopted by users, often in an organizational context that requires a collective adoption"* [5]. Unfortunately, it is practically impossible for the creators of the method to try out a method in numerous contexts who need to be trained in the domain of application [3], or designers in the domain of application who need to be trained on the method [9]. Moreover, no consensus exists on an agreed-upon set of indicators to use for comparing design methods [5]. Therefore, many papers do not provide objective evidence that the method will have the effect it purports under the conditions of prescribed use [5]. For instance, as reported by Vermaas [10]: "*The coexistence of these different traditions is now hampering further developments and usages of functional description in academia and industry*" [10].

Due to these barriers and many others – e.g., the publish-or-perish culture [3] –, too many publications do not include a validation [3,11] of the presented work. This not only lowers the intrinsic value of our scientific research but also undermines our recognition in the industry [3] since professional practitioners cannot determine which method or tool is better to employ in a specific context [9].

### *1.2. Problem*

Most research studies cited in the introduction agree upon the fact that, as stated in [12], *"Design research is increasingly weak in comparison with other fields; without action to increase scientific, theoretical, and methodological rigour there is a real possibility of the field being superseded and becoming obsolete through lack of impact.*".

Thus, when a researcher is willing to compare a set of solutions addressing the same design problem, he defines metrics to compare the candidate solutions with or without a systematic use case dataset. Doing this sort of benchmarking – e.g. in Model-Based Systems Engineering [13,14] – clarifies the current state of the art as it helps to identify the best solutions – in absolute or for a given context –, but the difficulties to collect and access the candidate solutions, the lack of standardized measures of performances and datasets, and the use without the creators being actively involved makes the benchmarking results too subjective.

In addition to the assessment of the practical challenges, few authors worked out methodological guidelines – e.g. the validation square [15], the Design Research Method (DRM) [6], the spiral of applied research [8], validation techniques from medicine [9] – to improve the quality of research in engineering design, including the validation of the contribution. In practice, unfortunately, researchers too rarely follow such guidelines and do not encourage enough PhD students to spend time studying and following them. Therefore, it is still necessary to define the means to test, validate or refute the results that engineering design research produces [5,6,16]. Thus, considering the existing methodological guidelines as well as the numerous practical challenges that are summed up in the introduction, the following research question emerged:

**RESEARCH QUESTION:**

How can we, in practice, help researchers in engineering design to methodologically improve the validation of their work?

### *1.3. Contribution*

By looking at research fields related to engineering design – e.g. combinatorial logic[1], machine learning[2], natural language processing[3], transformation tool[4], SAT solving[5], etc. – Benchmarking environments are gaining more and more interest as means to evaluate research results. Industrial challenges and hackathons are even open to everyone via web-based platforms, such as Github or Kaggle as well as websites of organizations (e.g. NASA[6]).

Although we often hear – with a grain of truth – that results of research in computer science are easier to benchmark, some scattered initiatives aim at developing elements of benchmarks in engineering design too, such as:

- **design problems** open to candidate design methods and tools – e.g. the NASA multiphysics optical instrument design problem[7],
- **standardised datasets** for assessing comparable results – e.g. the Common Research Model in CFD[8], the landing gear system case study [17], additive manufacturing test artefact[9], the wheel brake system from SAE ARP 4761,
- **competitions** for ranking candidate solutions to a problem in engineering design – e.g. CAD shape retrieval contests[10].
- **cartographies** of candidate solutions – e.g. the OMG and INCOSE's cartography of MBSE methods[11].

Nevertheless, as reported in the field of model-based design [18], to compare research proposals, a community platform with common goals, hypotheses, metrics for improvements, datasets, etc. is still missing. The need for a platform in which academics can exchange use cases with practitioners in the industry was also reported in [3].

[1] github.com/lsils/benchmarks
[2] reproducibility-challenge.github.io/iclr_2019/
[3] ixa2.si.ehu.eus/stswiki/index.php/Main_Page
[4] www.transformation-tool-contest.eu/
[5] www.satcompetition.org/
[6] www.nasa.gov/aeroresearch/resources/design-competitions-challenges
[7] github.com/nasa-jpl/design-challenge-multiphysics
[8] commonresearchmodel.larc.nasa.gov/2012/01/19/hello-world-2/
[9] www.nist.gov/topics/additive-manufacturing/resources/additive-manufacturing-test-artifact
[10] www.shrec.net/
[11] www.omgwiki.org/MBSE/doku.php?id=mbse:methodology

Thus, one year ago, a French national open science initiative has been launched to develop a community-based environment where researchers in engineering design can collaboratively build an ecosystem of design methods and tools that facilitates their benchmarking, that is, "*the process for rapidly learning the essence of a desired field of knowledge with an interest in finding its leading-edge"*. This paper report on the first six months of the national open science initiative and propose:

**RESEARCH PROPOSAL:**

1. A set of guidelines that helps researchers to define the Minimum Viable Content of a benchmark, and
2. An implementation of the guidelines in an online version-control and community-based platform.

Section 2 reviews the literature to elucidate a functional specification of the benchmarking platform. Section 3 proposes a set of guidelines to support researchers in the definition of the Minimum Viable Content of a bench. Finally, section 4 discusses the implementation of the guidelines in a community-based open science benchmarking environment.

## 2. Literature review

The introduction already reviewed and referred to numerous existing works to show that engineering design research suffers from a lack of validation rigour. This section goes through each paper in this bibliography to capture the needs that the benchmarking platform shall satisfy and to classify them into labelled topics. Note that the needs are expressed at different levels, but they will be clarified and articulated in a tree in Fig.1.

According to the literature, the benchmarking platform shall enable researchers in engineering design to:

**Define research**
- Consider qualitative and/or quantitative research studies [4]
- Consider multi-disciplinary research studies [8]
- Consider research that aims at understanding design [8]
- Consider research that aims at improving design [8]
- Clearly state the expected effects of a method [3]

**Collect/Share data**
- Build empirical databases to store qualitative and quantitative results from empirical design research [5]
- Accumulate a sufficient number of similar studies [5]
- Make the state-of-the-art of a research field publicly available [18]
- Dynamically collect new methods and tools [3]

**Compare results**
- Agree upon hypothesis across studies [5]
- Agree upon a set of (quantitative) measures for comparing results [5]
- Benchmarks are not selected by the tool developers [18]
- Provide a fair comparison of tools in an environment equal to all participants [3,18]
- Compare the power of novel approaches [18]
- Provide an exhaustive comparison of tools by covering the whole spectrum of interesting test cases [18]
- Characterise the properties of the application case [3]

**Provide objective evidence of progress**
- Define the theoretical contribution to knowledge [4]
- Define the empirical contribution to practice [3,4]
- Conduct meta-analysis [5]
- Capture causal influences such as applying a design method or tool makes it more likely that something is achieved [3,5]
- To measure the effects claimed [3]
- Increase scientific rigour [5]
- Give industry partners the confidence that the methods and tools are ready to be deployed and add value to their operations [3]
- Document the outcome of benchmarking such that the experiments performed are repeatable [18]
- Show where progress has been made and derive new research questions from results and then create new benchmarks that cannot be handled by current solutions and hand them over to the community to get solutions [18]
- Produce usable results with limited time and budget [3,8]

**Build an engaging ecosystem**
- Collaborative space [5]
- Gathers within groups of people with different skills and inclinations [3]
- Community-organised evaluation [18]
- Share a common terminology [3]
- Build work on each other [3]
- Articulate differences [3]
- Know "what works, for whom, under which circumstances" [5]
- Identify design problems as benchmark design problems, and then allocate different design methods or tools, including the specific design method to be studied and a comparable tool. [5]
- Select design methods and tools with ease depending on what is being designed [3]
- Pick up an existing method or tool to improve it [3]

## 3. Proposal

This section progressively goes from needs to elements of the proposed benchmark environment solution. First, we will distil a functional specification of the benchmarking platform based on the combination of needs identified in the literature and the ones captured during a workshop. A means-ends network is used to hierarchically organise the needs. Second, based on the functional specification, we propose a set of guidelines to support researchers in the definition of the Minimum Viable Content of a bench. Third, we detail how to implement the guidelines in a community-based open science benchmarking platform.

### *3.1. Functional specification*

As this issue requires a serious and collective effort, in January 2021, a two-hour webinar was organized to bring the validation problem up to the S.mart French national special interest group (SIG) of researchers in industrial engineering

(slides[12], and replay[13] are in open access). Among the 30 researchers and PhD students who attended the event, 14 decided to engage in the specification and development of the community-based platform. During the webinar, a document was accessible online and anyone was free to comment anonymously (an extract of the notes[14] is in open access). Based on the talk and the free notes, we identified two fundamental goals (benchmarking and sharing) and derived functions that the benchmarking platform could meet:

- G1. Comparing design methods and tools
  - To cartography proposals
  - To define datasets
  - To develop an open science space
  - To make sure results are reproducible
  - To standardize the types of objects we need to validate
  - To standardize the measures of performances
  - To create challenges
- G2. Sharing knowledge with peers in a community
  - To share bibliographies in open access
  - To share design papers with an exemplary validation
  - To share software in open access
  - To share datasets in open access
  - To share validation practices in a community

Two months later, in March 2021, a workshop occurs during the general assembly of the SIG S.mart where two groups in parallel spent 45 minutes in a virtual room to co-design a one-year roadmap based on the outputs of the first workshop. Finally, both groups joined in a single virtual room for consolidating both roadmaps into a single one (roadmaps[15] are in open access). During this workshop, two important needs emerged: 1) to clarify the taxonomy of the objects to validate, and 2) to build benchmarks.

To be able to build this benchmarking environment, we merge the needs found in the literature and the needs elucidated ding the workshops, and we organize them in a means-ends network that articulates "means" objectives contributing to the achievement of the fundamental "ends" objectives (see Fig. 1).

### *3.2. The Minimum Viable Content of a benchmark*

In practice, on the benchmarking platform, when a researcher desires to create a new benchmark to conduct an exhaustive and rigorous validation with a community of experts who will potentially join him in the adventure, he faces difficulties knowing which content he should define. Thus, as it exists the Minimum Viable Product in agile approaches, a proposition of a Minimum Viable Content (MVC) of a benchmark that satisfies the needs enumerated in Fig 1 is presented. Note that the MVC is different from a template as it concerns the storage of data and documentation. Moreover, it does not prescribe the form of the documentation but concentrates on the content.

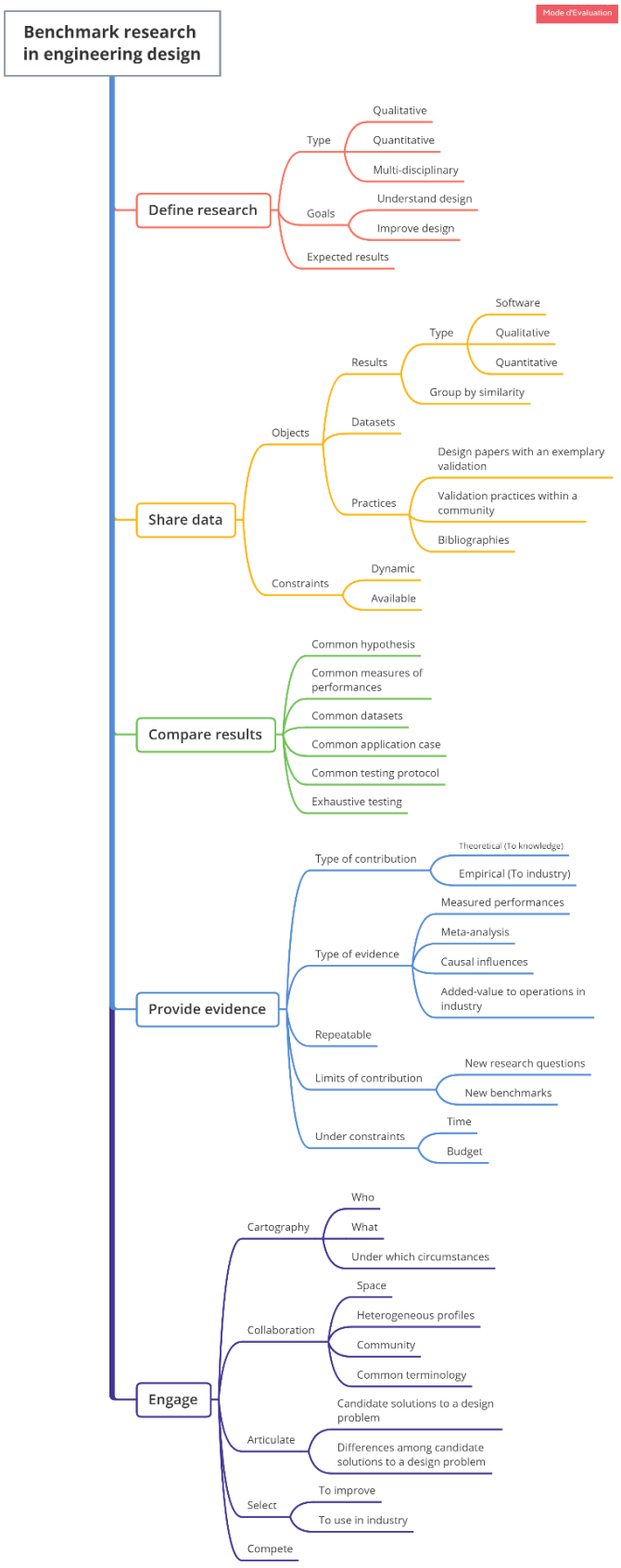


Fig. 1. Needs that the platform shall meet

[12] github.com/GIS-S-mart/Welcome/blob/main/Papers/20210324-Slides.pdf

[13] videos.univ-grenoble-alpes.fr/video/21692-webinar-smart-validation-des-recherches-en-conception/

[14] github.com/GIS-S-mart/Welcome/blob/main/Papers/20210324-Webinar_smart_validation_framapad.pdf

[15] miro.com/app/board/o9J_lOfryN8=/?invite_link_id=417150533172

**Introduction**: the documentation of the benchmark shall contain an introduction section that provides the gist of the benchmark. Thus, it shall define the type of research you are conducting (qualitative and/or quantitative), the main discipline(s) it contributes to, the goal of the study (understand or improve the design body of knowledge), the type of contribution (theoretical or empirical), objects you plan to validate (theory, method, tool, process), and the expected results.

**Glossary:** the benchmark document shall contain a glossary that enumerates the keywords and clarify their meanings with a definition and potential synonyms to avoid ambiguities within the community.

**Goals:** the documentation of the benchmark shall explicitly state the studied object and the pursued goals.

**Collection of similar solutions:** the benchmark shall collect qualitative and/or qualitative data supporting the similar solutions that have been submitted by the members of the community. The documentation of the benchmark shall clearly state the hypothesis upon which the solutions were built. The set of competing solutions is not fixed but follows the dynamics of the community and the data are in open access. Since open data are visible to the community, it encourages researchers to increase the academic rigour of their work.

**Comparison criteria:** the documentation of the benchmark shall include the criteria (quantitative if possible) that support the comparison of similar solutions submitted to the benchmark. The definition of the criteria is a collaborative task that the experts shall achieve together in synchronous or asynchronous meetings to get an agreed-upon reference. Like similar solutions, the list of criteria is not a fixed set but evolves according to suggestions made by experts in the community or new candidate solutions.

**Datasets**: the benchmark shall collect all the datasets that were used by the researchers in their evaluations. By enabling everyone to evaluate its solution on datasets that have been produced by peers, obtaining a fair comparison of solutions. Datasets shall be in open access to facilitate reuse and consequently reduce the cost and time of the evaluation. The number and content of datasets for a benchmark are not fixed but evolve with the dynamic of the community.

**Comparison of solutions:** the documentation of the benchmark shall report the results of the up-to-date comparison of the similar solutions that have been submitted to the benchmark. The comparison shall make sure that the results were obtained with a common testing hypothesis, comparison criteria, dataset, application case, and testing protocol. In addition, to compare comparable solutions, each solution shall be evaluated on a maximum of datasets to cover the whole spectrum of interesting tests.

**Evidence of progress:** the documentation of the benchmark shall provide evidence of progress, such as a series of improvements of a quantitative measure of performance, a meta-analysis, or a causal graph – e.g. an influence diagram [6] – to capture the impacts of the results on more fundamental objectives as well as the subsequent enhancements achieved by the community. The documentation of the benchmark shall also state the limits of the solutions submitted to the benchmark to define new research questions and new benchmarks in turn.

**References**: the documentation of the benchmark shall list the main papers related to the solutions submitted to the benchmark and potentially the datasets.

### *3.3. A community-based open science benchmarking platform*

A community-based open science benchmarking platform (www.github.com/GIS-S-mart) implements this Minimum Viable Content of a benchmark and provides a collaborative space for researchers in engineering design,

***What is the benchmarking platform?*** The platform uses Github rather than alternatives – e.g. Gitlab – to provide access to everyone without depending on remote administrators. Thus, the platform is a set of Github repositories hosted under the organization S.mart. The main repository stands for the welcome page (www.github.com/GIS-S-mart/Welcome) that gathers all the details (motivations, code of conduct, news, contribution process, list of existing benchmarks, project team, etc.) of the benchmark platform. Then, there is a repository named "Benchmark-template" (www.github.com/GIS-S-mart/Benchmark-Template) that is an implementation of the Minimum Viable Content in a Markdown page that anyone can download and reuse as a base for the documentation of his new benchmark. The rest of the repositories are the first benchmarks that were created by researchers and consequently have a more or less advanced level of progress. More advanced ones are public, whereas work in progress is kept private.

***What is a benchmark?*** A benchmark is a version-controlled Github repository that contains a Markdown file to document the benchmark, cloud storage for storing data, and a community. Markdown is a very simple file format that can be edited with WYSIWYG editors. The cloud storage supports any format, but it is more adapted to software.

***Where to store the data?*** If a researcher wants to share large datasets - e.g. 3D point clouds -, then he uses a dedicated repository (private disk servers, cloud solutions such as Google Drive, etc.) and implements hyperlinks into the Markdown documentation to provide access to it. Any data stored in Github or dedicated open data solutions (e.g. Zenodo, Harvard Dataverse, or Software heritage) can be identified with a DOI for future citations.

***How to join the community?*** Anyone with a Github account can join the community of a benchmark. Joining the community can be achieved by several means. One can browse the list of open issues to contribute to the existing benchmark by submitting a new solution, a new dataset, new criteria of comparison, or modifying existing ones. He may also simply create a new benchmark with the hope to create a new community (see Fig. 2). Rather than open a new issue that should lead to changes, one can open a new discussion (see Fig. 3) that will enable participants to engage in a new reflection (adding a new criterion or dataset) with the need to collect a maximum of feedback before opening a new issue for requesting a change.


Fig. 2. Someone suggesting existing benches in additive manufacturing

***When to create a new benchmark?*** Before all, to avoid the mushrooming of irrelevant repositories as it occurs with papers, a new benchmark shall be created after a consensus emerged from a discussion within the community. The opening of a discussion for creating a new benchmark should be motivated by new goals requiring the validation of new methods and tool that aims to improve or understand the design activity. A benchmark should be neither too broad to avoid the comparison of non-comparable solutions nor too specific to avoid isolated knowledge islands [3] with a single participant like it often happens in the open-source movement.

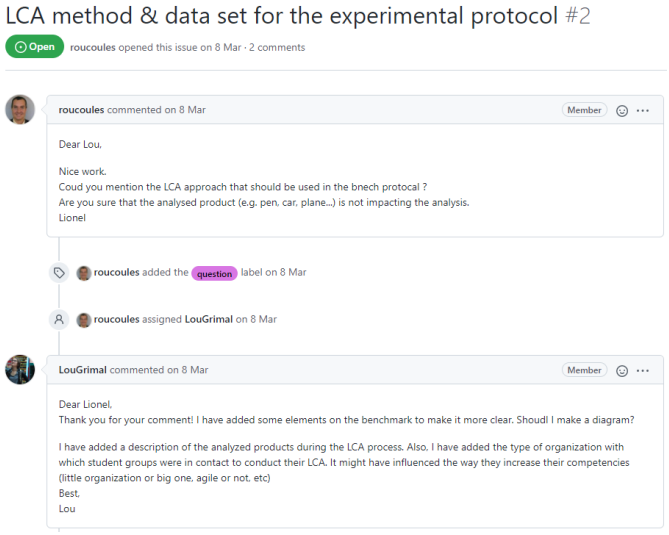


Fig. 3. An open issue to suggest improvements to an LCA bench

## 4. Conclusion

For fifteen years, it emerges a consensus: "*There is this concern that design research does not live up to the standards of science: it is creating in a sense too many theories and models, which jeopardises the coherence of the discipline and which indicates that design research does not yet have the means to test and refute design theories and models*" [16]. Researchers in engineering design worked out methodological tools to improve research quality, especially the validation. Unfortunately, those tools are rarely adopted.

To increase the quality of design research, the authors propose an online open-source community-based benchmarking platform where researchers can engage in the building of a sustainable validation ecosystem. In this paper, they identify the needs that the platform shall satisfy, propose guidelines to help researchers to define the Minimum Viable Content of a benchmark, and detail a Github implementation.

Future work will concentrate on the continuous improvement of the current platform (benchmarks, instructions, workflows…) based on the active members of the special interest group before inviting other communities.

## Acknowledgements

We thank all members of the French S.mart special interest group who contributed to the discussions.

## References

[1] Tomiyama T. How to write a good STC Dn paper. unpublished article, presented at the CIRP STC Dn meeting in February 2021.
[2] Cantamessa M. An empirical perspective upon design research. J Eng Des 2003;14:1–15.
[3] Gericke K, Eckert C, Campean F, Clarkson PJ, Flening E, Isaksson O, et al. Supporting designers: moving from method menagerie to method ecosystem. Des Sci 2020;6:e21.
[4] Isaksson O, Eckert C, Panarotto M, Malmqvist J. You need to focus to validate. Proc Des Soc Des Conf 2020;1:31–40.
[5] Hein AM, Lamé G. Evaluating engineering design methods: taking inspiration from software engineering and the health sciences. Proc Des Soc Des Conf 2020;1:1901–10.
[6] Blessing LTM, Chakrabarti A. DRM, a Design Research Methodology. London: Springer London; 2009.
[7] Gericke K, Meissner M, Paetzold K. Understanding the context of product development. Proc. 19th Int. Conf. Eng. Des. Des. Harmon. Vol. 3 Des. Organ. Manag., Seoul, Korea: 2003.
[8] Eckert C, Clarkson PJ, Stacey. The spiral of applied research: A methodological view on integrated design research. Proc. 14th Int. Conf. Eng. Des., Stockholm, Sweden: 2003.
[9] Frey DD, Dym CL. Validation of design methods: lessons from medicine. Res Eng Des 2006;17:45–57.
[10] Vermaas PE. The coexistence of engineering meanings of function: Four responses and their methodological implications. Artif Intell Eng Des Anal Manuf 2013;27:191–202.
[11] Barth A, Caillaud E, Rose B. How to validate research in engineering design? 18th Int. Conf. Eng. Des. (ICED 11), Impacting Soc. through Eng. Des. , Vol. 2 Des. Theory Res. Methodol., Copenhagen, Denmark: 2011, p. 41–50.
[12] Cash PJ. Developing theory-driven design research. Des Stud 2018;56:84–119.
[13] Weilkiens T, Scheithauer A, Di Maio M, Klusmann N. Evaluating and comparing MBSE methodologies for practitioners. 2016 IEEE Int. Symp. Syst. Eng., IEEE; 2016, p. 1–8.
[14] Di Maio M, Weilkiens T, Hussein O, Aboushama M, Javid I, Beyerlein S, et al. Evaluating MBSE Methodologies Using the FEMMP Framework. 2021 IEEE Int. Symp. Syst. Eng., IEEE; 2021, p. 1–8.
[15] Carolyn C. Seepersad, Kjartan Pedersen Jan Emblemsvåg, Reid Bailey, Janet K. Allen FM. The Validation Square: How Does One Verify and Validate a Design Method? Decis. Mak. Eng. Des., ASME Press; 2006, p. 303–13.
[16] Vermaas PE. Design Theories, Models and Their Testing: On the Scientific Status of Design Research. An Anthol. Theor. Model. Des., London: Springer London; 2014, p. 47–66.
[17] Boniol F, Wiels V, Aït-Ameur Y, Schewe K-D. The landing gear case study: challenges and experiments. Int J Softw Tools Technol Transf 2017;19:133–40.
[18] Dubois C, Famelis M, Gogolla M, Nobrega L, Seidl M, Völter M, et al. Research Questions for Validation and Verification in the Context of Model-Based Engineering. Int. Work. Model Driven Eng. Verif. Valid. - MoDeVVA, Miami, United States: 2013, p. 67–76.