CHAPTER 1

How to measure diversity actionably in technology

Authors: Md Montaser Hamid, Amreeta Chatterjee, Mariam Guizani, Andrew Anderson, Fatima Moussaoui, Sarah Yang, Isaac Tijerina Escobar, Anita Sarma, and Margaret Burnett

How do you measure technology's support for diverse populations in a way that is actionable, and can lead to more inclusive designs of the technology? This chapter presents a method and the validated GenderMag survey that powers the method. The survey measures diversity gaps in technology in a fine-grained way, and the method shows how to use it to translate an empirical study's findings into actionable design directions.

Introduction

> *Measurement is the first step that leads to … improvement.* (IBM quality expert H. James Harrington) [11]

Many scientists and researchers, including us, agree with Mr. Harrington. When considering diversity, a reason for measurements is often a desire to change something to improve the support for diversity.

Our interest lies in measuring the diversity of a user population that a software system intends to support. Improving how well a software system supports diverse users in technology requires diversity measurements that are truly actionable—not just a demographic measurement (e.g., “we don’t support women as well as other people”; or “only 37% of women would recommend our software, compared to 51% of other people”). Demography-based measurements can point out what features disproportionately affect diverse users and how often these issues arise but are incapable of explaining why these issues exist in the first place. Those why’s are the missing link that enable translating the empirical study findings into actionable design fixes.

To obtain those missing why’s, what is needed is a fine-grained measurement device that relates technology misfires with diverse individuals’ traits relevant to the usage of technology. Toward that end, we have developed a diversity measurement method based on the GenderMag facets enabled by a GenderMag facet survey. The GenderMag facets represent different cognitive styles that impact how individuals go about using technology, in which the differences (statistically) cluster by gender. The GenderMag facet survey provides a new, fine-grained method for understanding diversity gaps in technology and in technology-related artifacts (e.g., user interfaces, documentation, user manuals). Although our previous work used facet surveys, this is the first time we explain the exact steps of the scoring and the validation process. The survey enables: (1) extracting information on who runs across which inclusivity bugs and why; (2) comparisons between a technology’s before/after diversity support; and (3) developers and designers to understand how to make the empirical results *actionable*.

Background: The GenderMag facets

The GenderMag facet survey is a companion to the GenderMag method [2]. GenderMag is an evidence-based inclusivity evaluation method that software practitioners can use to find and fix inclusivity bugs.

GenderMag has been used for a wide range of applications [4, 8, 9, 10, 12, 14, 15, 17].

At the core of GenderMag are five problem-solving styles called facets in GenderMag (Figure 1-1), each of which is backed by extensive foundational research [2, 16], and has a range of possible values. A few values within each facet's range are brought to life by the three GenderMag personas: "Abi", "Pat", and "Tim." Statistically, Abi's facets are more common among women and Tim's are among men, whereas Pat has a mix of Abi's/Tim's facets plus a few unique ones.

Each facet describes how different genders approach problem solving when using technology. For example, women may have a process-oriented learning style like Abi which means they would prefer to learn new technologies in the context of a tutorial or an explicit process. When looking for information to progress, some genders may be more selective (Tim's processing style) as in they pick the first promising option instead of reading through all the information. These facets can help designers pinpoint how to better support all genders within technology and the facet survey enables them to measure a respondent's facet values (Figure 1-1).

The facet survey: What it is

The GenderMag facet survey (Figure 1-2) is a validated Likert-scale survey that collects a respondent's particular facet values for each of the five facets in Figure 1-1. We initially created it as part of a longitudinal field study at Microsoft that occurred in 2015-2016 [3].

At that time, Microsoft had just developed a strong interest in supporting diversity and inclusion within its *products*—not just its workforce climate. This timeframe coincided with the emergence of GenderMag and Burnett's sabbatical at Microsoft, and subgroups of Microsoft employees were considering using all or portions of it. However, GenderMag's generality and applicability to their products had not been established yet, and some employees wondered whether the facet distributions across genders that the GenderMag team had seen

elsewhere really applied to their customers.

Self-Efficacy *Abi*: Lower self-efficacy than their peers about unfamiliar computing tasks. If tech problems arise, often blames self and might give up as a result. *Tim*: Higher self-efficacy than their peers with technology. If tech problems arise, usually blames the technology. Sometimes tries numerous approaches before giving up. *Pat*: Medium self-efficacy with technology. If tech problems arise, keeps on trying for quite a while.
Motivations Uses technology... *Abi*: Only as needed for the task at hand. Prefers familiar and comfortable features to keep focused on the primary task. *Tim*: To learn what the newest features can help accomplish. *Pat*: Like Abi in some situations and like Tim in others.
Learning Style *Abi*: Learns best through process-oriented learning; (e.g., processes/algorithms, not just individual features). *Tim*: Learns by tinkering (i.e., trying out new features), but sometimes tinkers addictively and gets distracted. *Pat*: Learns by trying out new features, but does so mindfully, reflecting on each step.
Information Processing *Abi* and *Pat*: Gather and read everything comprehensively before acting on the information. *Tim*: Pursues the first relevant option, backtracking if needed.
Attitude Toward Risk *Abi* and *Pat*: Risk-averse, little spare time; like familiar features because these are predictable about the benefits and costs of using them. *Tim*: Risk tolerant; ok with exploring new features, and sometimes enjoys it.

Figure 1-1: The GenderMag facet types and their values for each persona [2]. The colors are used throughout this chapter to associate the survey questions/scoring with these facets.

Microsoft's "Team C2" was the first to raise this question, and to answer it, they sought validation within their product's customers. Thus, they collaborated with the GenderMag team to develop and run a GenderMag facet survey which is framed within the GenderMag method. This survey helped them in validating the GenderMag facet values and accompanying gender distributions in their customer base. The survey results also answered the question that Team C2 had sought to answer: whether the GenderMag facets were indeed pertinent to their own customers.

1. I am able to use <insert type of technology> when…
   a. I have just the build-in help for assistance.
   b. I have seen someone else using it before trying it myself
   c. no one is around to help me if I need it.
   d. someone else has helped me get started.
   e. someone has shown me how to do it first.
   f. I have used similar technology before to do the same task.
   g. I have never used anything like it before.
2. I am not confident about my ability to use and learn <insert type of technology>. I have other strengths.
3. I make time to explore <insert type of technology> that is not critical to my job.
4. One reason I spend time and money on <insert type of technology> is because it's a way for me to look good with peers.
5. It's fun to try new <insert type of technology> that is not yet available to everyone, such as being a participant in beta programs to test unfinished <insert type of technology>.
6. I enjoy finding the lesser-known features and capabilities of the <insert type of technology> I use.
7. I explore areas of <insert type of technology> before it is time for me to use it.
8. I'm never satisfied with the default settings for my <insert type of technology>; I customize.
9. I want to get things right the first time, so before I decide how to take action, I gather as much information as I can.
10. I always do extensive research and comparison shopping before making important purchases.
11. When a decision needs to be made, it is important to me to gather relevant details before deciding, in order to be sure of the direction we are heading.
12. I avoid "advanced" buttons or sections in <insert type of technology>
13. I avoid activities that are dangerous or risky.
14. Despite the risks, I use features in <insert type of technology> that haven't been proven to work.

Completely Agree			Neither Agree nor Disagree			Completely Disagree		
1	2	3	4	5	6	7	8	9

Figure 1-2: The facet survey. (Top) Question colors indicate the facet being measured (Figure 1-1). (Bottom): All questions use a 9-point Likert scale.

More importantly, the survey results revealed a measurement benefit we hadn't anticipated: it offered a measure of diversity outcomes at a higher resolution than standard demographic measures. In this chapter, we define a *higher resolution measure* as one that can discriminate between two points that, with a lower-resolution measure, cannot be discriminated. Applying this concept to diversity outcome measurements, suppose that 67% of women run into barriers with a

particular feature of a technology product, and 33% do not. What are the differences between someone in the 67% and someone in the 33% group other than the outcome? If all we have is their gender demographics, all we know is that they are women, and we cannot see their differences. However, if we also have their facet values, we can see differences that gender demographics alone cannot reveal.

The facet survey can be used in several ways, but the primary use we discuss in this paper is to obtain fine-grained measurements of diversity in an empirical investigation.

Scoring the survey

After participants have responded to the facet survey, we can score their responses using the survey key in Figure 1-3. Since Abi and Tim are the personas who represent the endpoints of each facet's spectrum of possible values, we use those persona names to relate a participant's responses to these two endpoints. We score using the following steps:

Step 1 (*Complement*): Convert the answer scores to numbers from 1 (Completely disagree) to 9 (Completely agree). For some questions, closer to 9 is Tim-like. But the opposite is true for how Questions 2 and 9–13 are worded, so for these questions, "reverse" the participants' responses to their tens' complement. (i.e., convert "9" to "1", "8" to "2.")

Step 2 (*Sum each facet*): For each participant, sum the results of Step 1 for each facet. The colors in Figure 1-1 and 1-2 represent facets. This step results in 5 scores per participant, one score for each facet.

Step 3 (*Calculate facet medians*): The scores are not "absolute." Rather, they are *relative to a participant's peer group*. For example, a group of college students would be expected to have different level of computer self-efficacy, different style of learning technology, etc., than a group of retired people. To find the middle of the peer group (we assume a peer group is the participants recruited for the study), calculate the median "sum of scores" of all the participants from the same peer group, for each facet.

1a, 1b, 1c, 1d, 1e, 1f, 1g	High Self Efficacy (Tim)
2	Low Self Efficacy (Abi)
3, 4, 5	Motivations: Technology for its own sake (Tim)
6, 7, 8	Learning: Tinkerer (Tim)
9, 10, 11	Comprehensive Information Processing (Abi)
12, 13	Risk Averse (Abi)
14	NOT Risk Averse (Tim)

Figure 1-3: Facet survey key. The more strongly the participant agrees on a question, the closer their facet value is to the endpoint (persona name) shown.

Step 4 (*Tag each participant's facet score*): To the right of the median (above) is Tim-like, otherwise it is Abi-like. If the participant's facet score is the facet median, then it is up to you to decide whether they are an Abi or Tim. You can decide on this in a way that helps balance the sample sizes, or you can add a third tag (Pat-like).

In the end, each participant has a 5-tuple tag representing each facet. Most participants turn out to have a mix of facet values. For example, a participant might have self-efficacy, motivations, and a risk attitude closer to Abi's, but information processing style and learning style closer to Tim's. The scores calculated from the facet survey then can be used to analyze when a technology is failing to be inclusive, why it is so, and exactly who are affected by it, as we detail next.

From scores to understanding to actions

We show how to analyze these scores in a way that points toward fine-grained understanding and then actionably using Team V as a running example [17]. Team V had two versions of a prototype: The "Before" version was the one currently in production usage, and the "After" version was a redesign to fix 6 inclusivity bugs of the Before version that Team V had found using the GenderMag evaluation method.

To understand in a fine-grained way their inclusivity progress, equity progress, and where design actions were still needed, Team V empirically evaluated both versions in a between-subject user study, in which Team V's participants responded to the facet survey and then worked their way through the prototype's main use-cases.

Fine-grained diversity comparisons between versions: Team V used the facet survey to compare their Before vs After versions' participants' encounter with the inclusivity bugs. Figure 1-4 aggregates the results for all 6 inclusivity bugs by counting the facet responses of Team V's participants who faced the bugs. For example, if a Before participant had 3 Abi facets and 2 Tim facets, they would add 3 to the "Before" orange bar and 2 to the "Before" blue bar.

As the figure shows, in the Before version, Abi facets were more impacted by inclusivity bugs than Tim (34 Abi facets (●) vs. 26 Tim facets (■)). The After version reduced the facets impacted for both: Abi 13 (●) and Tim 17 (■). Therefore, the After version improved inclusivity for both Abi- and Tim-faceted users; but it was still not equitable since Tim-like facet values were more impacted than Abi-like facet values.

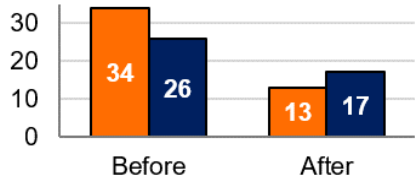

Figure 1-4: Number of observed facets in the facet survey responses of Team V's participants who faced inclusivity bugs. Orange: Abi facets. Blue: Tim facets.

Fine-grained understanding of who and why: To understand how to fix a bug, such as Bug#4 (Figure 1-5), we first need to know who experienced it and why. In this bug, six Team V's participants (Before1, Before2, Before4, Before5, Before8, and Before10) faced the inclusivity bug in the Before version and two (After7 and After10) in the After version. (The participants are ordered by the number of their Abi (●) vs. Tim (■) facet values, with Abi's at the top and Tim's at the bottom.)

In the Before version, six of Team V's participants spanning every facet value experienced Bug#4 difficulties, with 16 Abi-facet count (●) and 14 Tim-facet count (■). In the After version, only 2 of Team V's participants faced the bug, with the Abi-facet count (●) now down to 2 and the Tim-facet count (■) down to 8. This reduction means that the Bug#4 fixes brought inclusivity by improving the prototype for both Abi-

and Tim-like users. But the After version did not achieve equity, with Tim-like facet values facing more difficulties than Abi's (8 ■ vs 2 ●).

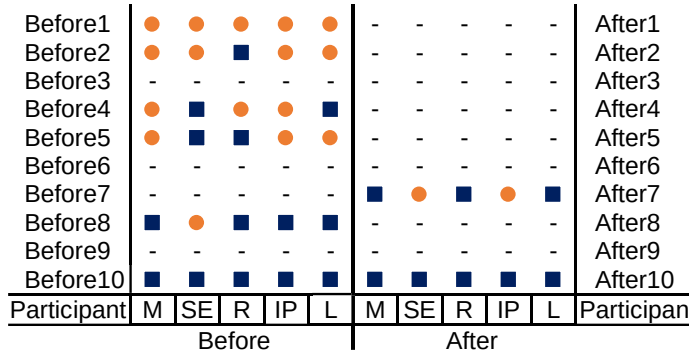

Figure 1-5: Results of Bug#4 in [17]. Facets of the 6 Before and 2 After versions' participants with action failures. -: participant had no action failures. M=Motivations; SE=Self-Efficacy; R=Risk; IP=Info Processing; L=Learning.

Where designers' actions helped and where more were needed: These counts showed that designers' Bug#4 remedies in the After version had been very successful: users with all 5 Abi-like facets and 4 Tim-like facets fared better than they had with the Before version. However, support for users with Tim-like *motivations* had not improved. This points designers directly toward designing further Bug#4 improvements to better support Tim's motivations (without sacrificing support for Abi's motivations); Guizani et al.'s "Why/Where/Fix" inclusivity debugging approach gives examples of how to do this [10].

To summarize, the Team V example shows how the facet survey can enable: fine-grained diversity comparisons between two versions; fine-grained understanding of "who" are being left out and their facet values; and designers where to take action by fixing inclusivity bugs based on the facet values of who's still being left out.

How we validated the survey

To validate the survey, we followed these steps below, but they were intertwined with each other and with the creation process.

Step 1:(pre-Validate) started with questions from other validated surveys;

Step 2: (Reliability) ran the survey and assessed response consistency using Cronbach alpha tests;
Step 3: (Cross-validate) cross analyzed results from administering to other populations, intertwined with Step 4;
Step 4: (Cluster analysis + Condense) cluster analyses to reduce the number of questions needed;
Step 5: (Demographic validation) quantitative comparison of facet responses with participants' gender identities; and
Step 6: (Empirical) a validation study comparing participants' survey responses with their verbalizations while working with the technology.

Much of this intertwined process was a joint effort with Microsoft's Team C2 [3]. In Step 1, we worked with Team C2 in a formative way, which we'll refer to as pre-Validation. In this step, we drew applicable existing questions from other validated surveys/questionnaires in the literature. This approach provided about two-thirds of the questions from established, validated questionnaires such as [6]. Although excerpting portions of a validated questionnaire cannot bring "validation" to the new questionnaire, the strong provenance of those excerpts enabled an evidence-based start to the survey. For facets with no validated questionnaire, we had to develop pertinent questions ourselves by drawing on existing research as much as possible. Stumpf et al.'s summary of gender-meets-technology literature covers much of the research base from which we drew [16].

After several iterative improvements, in a 2015 study, Team C2 ran the survey on 500 men and 500 women who were Microsoft's customers. For Step 2 (Reliability), we analyzed their results using Cronbach alpha tests [7], a widely used way of measuring the reliability of a set of questions. The results validated the survey's inter-item reliability. Specifically, the results were above the 0.8 level for two of the five facets (Information Processing Style and Self-Efficacy), above the 0.7 level for two others (Motivations and Risk), and at 0.691 for Learning Style. Cronbach alpha's above 0.8 is generally considered to be good and above 0.7 to be acceptable, but Churchill also argues that 0.6 should also be considered acceptable [5].

Word of the survey spread, and by Step 3 (Cross-validation), other interested teams began to work together to share and cross-analyze survey results. At the same time, (Step 4: Cluster analysis and condense), several Microsoft data scientists got involved to whittle down the number of questions needed by analyzing responses, so as to reduce possibility of survey fatigue. One team also validated their survey with user interviews and think-aloud studies.

One validity question that needed to be answered was whether the facet values reported by Team C2's 1000 survey respondents differed (quantitatively) by gender identity (Step 5: Demographic validation). One qualitative study in 2019 to answer this question involved 20 participants [17]; another in 2021 involved 1000 participants [1]. As Figure 1-6 shows, these participants' facet values did cluster by their gender where women skewed more towards Abi than men did.

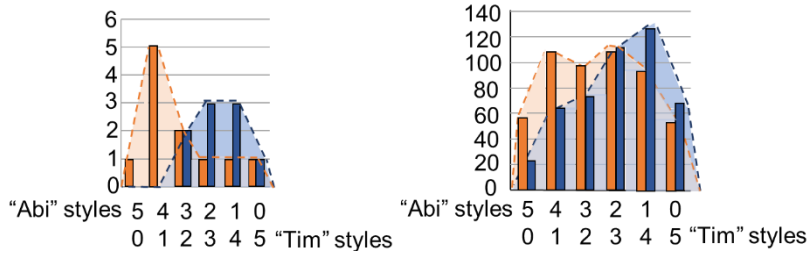

Figure 1-6. Facet survey results for 2 genders. x-axis: # facet values scored as (top row): "Abi"-like; (bottom row): "Tim"-like. y-axis: # participants. Example: the bar at "5 0" on the x-axis shows the number of participants with all 5 Abi-like facet values. (Left chart) [17], (Right chart) [1].

Finally (Step 6), in a 2022 think-aloud study [10], we compared participants' facet survey responses with their in-situ verbalizations during a think-aloud problem-solving task. Figure 1-7 shows results from this comparison. When an outline color (their in-situ verbalized facet value) is the same as the shape's fill color (their survey response), then their survey response matched that participant's verbalized facet value in that moment of their work. In total, 78% of participants' in-the-moment facet verbalizations aligned with their facet survey responses which suggest that the facet survey was a reasonable measure of participant's actual facet values.

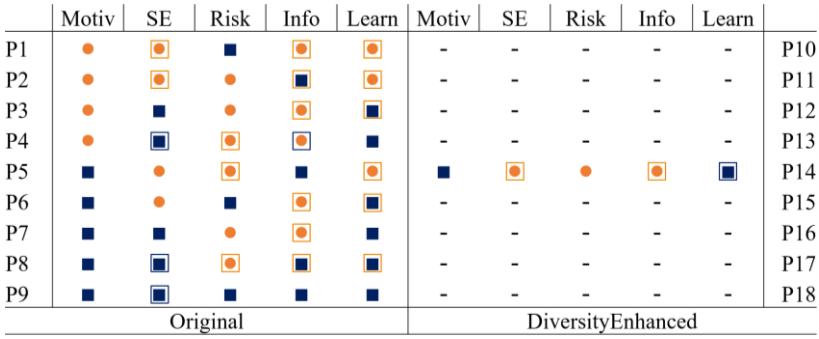

Figure 1-7: Think-aloud study participants [10] who ran into one set of inclusivity bugs with their facet values, validated with their in-situ responses. ● | ■ : the facet scores from the participants' survey responses for Abi-like and Tim- like facet values respectively; ◘ | ◘ : Abi-like | Tim-like facet values participants verbalized in-situ when they ran into a bug.

Key Takeaways

The key takeaways from this chapter are:

Fine-grained diversity measurements: The GenderMag facet survey measures technology's diversity gaps in a fine-grained way, showing not only who experiences which inclusivity bugs, but also *why* they experience each inclusivity bug they encounter.

Fine-grained comparisons: The survey enables comparisons between different prototype versions showing not only which version is more inclusive, but also why and for whom that version is more inclusive.

Actionable: The why's are *actionable*: designers can design fixes to an inclusivity bug around the facet values of those experiencing it.

Validated: The survey has been thoroughly validated.

The survey has uses beyond measurement, such as to select a facet-diverse set of participants [10], or for team building [13], but its main purpose is measuring diversity *actionably*. We invite researchers, developers, and designers everywhere to use it to gain new insights into both how to address their technology's diversity failures and how to repeat their technology's diversity successes.

Acknowledgments

We thank the many people who have helped to improve the survey. We thank Robin Counts, Hannah Hanson, and Ronette Lawrence. Robin initiated the idea of a facets survey and Ronette enabled Microsoft's support. Robin and Hannah not only contributed substantially to the research and development of its questions, but also to the administration and analysis of the survey's first trials. Thanks to Ronette, Microsoft permitted us to freely share the survey questions. This work has been supported in part by Microsoft; by National Science Foundation grants 1901031 and 2042324; and by USDA-NIFA/NSF grant 2021-67021-35344. Views expressed in this chapter are those of the authors, and not necessarily those of the sponsors.

References

[1] Anderson A, Li T, Vorvoreanu M, and Burnett M (2022) Diverse humans and Human-AI interaction: What cognitive style disaggregation reveals. https://arxiv.org/abs/2108.00588v3

[2] Burnett M, Stumpf S, Macbeth J, Makri S, Beckwith L, Kwan I, Peters A, and Jernigan W (2016) GenderMag: A method for evaluating software's gender inclusiveness. Interacting with Computers 28(6):760–787. https://doi.org/10.1093/iwc/iwv046

[3] Burnett M, Count R, Lawrence R, and Hanson H. (2017). Gender HCI and Microsoft: Highlights from a longitudinal study. IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), 139-143. 10.1109/VLHCC.2017.8103461.

[4] Chatterjee A, Letaw L, Garcia R, Reddy DU, Choudhuri R, Kumar SS, Morreale P, Sarma A, and Burnett M (2022) Inclusivity bugs in online courseware: A field study. 2022 ACM Conference on International Computing Education Research-Volume 1 (ICER

’22). ACM, New York, NY, USA, p 356–372. https://doi.org/10.1145/3501385.3543973

[5] Churchill GA (1979) A paradigm for developing better measures for marketing contrasts. Journal of Marketing Research 16(1):64–73. https://doi.org/10.2307/3150876

[6] Compeau DR and Higgins CA (1995) Computer self-efficacy: Development of a measure and initial test. MIS Quarterly 19(2):189-211. https://doi.org/10.2307/249688

[7] Cronbach LJ (1971) Test validation. In: Thorndike RL (ed) Educational measurement. American Council on Education, Washington DC, USA, p 443-507

[8] Cunningham SJ, Hinze A, and Nichols DM (2016) Supporting gender-neutral digital library creation: A case study using the GenderMag Toolkit. 2016 International Conference on Asian Digital Libraries (ICADL ’16). Springer, Cham, Switzerland, p 45–50. https://doi.org/10.1007/978-3-319-49304-6_6

[9] Gralha C, Goulão M, and Araújo J (2019) Analysing Gender Differences in Building Social Goal Models: A Quasi-experiment. 2019 IEEE International Requirements Engineering Conference (RE ’19). IEEE, New York, NY, USA, p 165–176. https://doi.org/10.1109/RE.2019.00027

[10] Guizani M, Steinmacher I, Emard J, Fallatah A, Burnett M, Sarma A (2022) How to debug inclusivity bugs? A debugging process with Information Architecture. 2022 IEEE/ACM International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS ’22). IEEE, New York, NY, USA, p 90–101. https://doi.org/10.1145/3510458.3513009

[11] Harrington HJ, Hoffherr GD, Reid RP, and Harrington R (1999) Area activity analysis. McGraw-Hill, New York, NY, USA

[12] Kanij T, Grundy J, McIntosh J, Sarma A, and Aniruddha G (2022) A new approach towards ensuring gender inclusive SE job advertisements. 2022 IEEE/ACM International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS

’22). IEEE, New York, NY, USA, p 1–11. https://doi.org/10.1145/3510458.3513016

[13] Letaw L, Garcia R, Garcia H, Perdriau C, and Burnett M (2021) Changing the online climate via the online students: Effects of three curricular interventions on online CS Students’ inclusivity. 2021 ACM Conference on International Computing Education Research (ICER ’21). ACM, New York, NY, USA, p 42–59. https://doi.org/10.1145/3446871.3469742

[14] Oleson A, Mendez C, Steine-Hanson Z, Hilderbrand C, Perdriau C, Burnett M, and Ko AJ (2018) Pedagogical content knowledge for teaching inclusive design. 2018 ACM Conference on International Computing Education Research (ICER ’18). ACM, New York, NY, USA, p 69–77. https://doi.org/10.1145/3230977.3230998

[15] Shekhar A and Marsden N (2018) Cognitive walkthrough of a learning management system with gendered personas. 2018 ACM Conference on Gender & IT (GenderIT ’18). ACM, New York, NY, USA, p 191–198. https://doi.org/10.1145/3196839.3196869

[16] Stumpf S, Peters A, Bardzell S, Burnett M, Busse D, Cauchard J, and Churchill E (2020) Gender-inclusive HCI research and design: A conceptual review. Foundations and Trends in Human-Computer Interaction 13(1):1–69. http://doi.org/10.1561/1100000056

[17] Vorvoreanu M, Zhang L, Huang Y-H, Hilderbrand C, Steine-Hanson Z, and Burnett M (2019) From gender biases to gender-inclusive design: An empirical investigation. 2019 SIGCHI Conference on Human Factors in Computing Systems (CHI ’19). ACM, New York, NY, USA, p 1–14. https://doi.org/10.1145/3290605.3300283