The Reliability and Validity of the ARIS Broader Impacts Rubric

Ellen R. Iverson, Kristin O'Connell, Janice D. McDonnell, Susan D. Renoe, and Liesl Hotaling

Abstract

More than 27 years have passed since the National Science Board identified Broader Impacts (BI) as one of two merit criteria for National Science Foundation proposals. Yet many researchers remain less certain of how to develop, implement, and assess a BI plan. This multimethod study of a BI rubric analyzed expert panels that included BI professionals and researchers for both content validity and reliability. Focus groups with researchers explicated both challenges related to BI plans and the potential value of the rubric. They revealed the challenges researchers have weighing proven strategies versus innovative strategies, a bias other scholars have documented. Researchers stated concern with how to weigh the different facets of the rubric to arrive at a single score. Moreover, researchers reported that their disciplinary fields influenced how they interpreted the audiences whose needs and interests may be met through BI plans. These distinctions represent a range of different types of community-engaged scholarship (e.g., public information network, community–campus partnerships, K–12 school partnerships). Finally, researchers found the BI rubric useful in evaluating and developing their own BI plans as well as their role in panels to ultimately strengthen the field of funded BI work.

For more than 25 years, panel reviewers have assessed proposals through the National Science Foundation (NSF) using *Broader Impacts* and *Intellectual Merit* as the two required merit criteria for all programs. In outlining the Broader Impacts (BI) criterion, the 1997 National Science Board (NSB) task force included suggested questions for the BI criterion that underscored areas relevant to community-engaged scholarship, such as "promoting teaching, training, and learning," "broaden[ing] the participation of underrepresented groups," or "enhance[ing] the infrastructure for research and education, such as … partnerships" (NSB, 1997). Since then, the NSB has affirmed the two criteria and, as part of its Vision 2030 report, recommended that the BI criterion be evaluated to better meet societal needs (NSB, 2011, 2020).

Despite the tenure of the BI criterion and the growth in robust community-engaged scholarship, BI has not received the systems and infrastructure of support that the Intellectual Merit criterion has. Intellectual Merit remains the backbone for advancement in higher education, with structures and systems for publication, promotion, and tenure legitimized by universities and professional societies; less policy recognizes community-engaged scholarship (Doberneck, 2016). The lack of guidance and clarity for the BI criterion was recognized by Congress, which identified eight specific goals for BI[1] and called for training and infrastructure for programs to help researchers implement BI plans (America Competes Reauthorization Act, 2010). Since then, the NSB (2011, 2020) and the NSF (2024) have stressed the importance of research benefiting

[1] The NSF states nine societal outcomes as examples for BI and notes that outcomes are not limited to only these. The outcomes include the eight identified by Congress and also *Infrastructure* as a ninth example. The set of nine consist of *Inclusion* (increasing and including the participation of women, persons with disabilities, and underrepresented minorities in STEM), *STEM Education* (improving education and educator development—at any level—in science, technology, engineering, and mathematics), *Public Engagement* (increasing public scientific literacy and public engagement with STEM), *Societal Well-being* (improving the well-being of individuals in society), *STEM Workforce* (developing a more diverse, globally competitive STEM workforce), *Partnerships* (building partnerships among academia, industry, and others), *National Security* (improving national security), *Economic Competitiveness* (increasing the economic competitiveness of the United States), and *Infrastructure* (enhancing infrastructure for research and education).

Note. This article is included in a special issue focused on the Implementation and Evaluation of the ARIS Broader Impacts Toolkit project, which is designed to advance the understanding of mechanisms and supports needed to develop effective Broader Impacts (BI) statements. The full issue can be found at https://jces.ua.edu/37/volume/17/issue/2

society and upheld it as a criterion for funding across programs.

A national grassroots effort led by BI professionals and higher education offices serving in this capacity recognized the need for training and resources. We define a BI professional as a specialist in an academic, nonprofit, government, private sector, or community-based organization who bridges the gap between scientific research and its potential benefits to society. Although these interdisciplinary experts often do not identify themselves as BI professionals, they work to ensure that scientific research serves the public good in a variety of ways, including but not limited to fostering public engagement, enhancing education, promoting diversity and inclusion, and contributing to economic development. BI professionals' roles vary widely and may support this work during the conceptual pre-award phase through engagement with partners during and after funded programming.

This network of BI professionals came together through the support of NSF grants to develop a community of practice, first as the National Alliance of Broader Impacts (NABI) and with subsequent funding as the Center for Advancing Research Impacts in Society (ARIS). These communities developed and shared tools such as the guiding principles for BI (ARIS, 2020; NABI, 2015). It is from this need that a set of BI professionals (including two of the authors) with expertise in developing and implementing BI plans associated with NSF awards developed the initial BI rubric in 2020. The initial set of 13 criteria was rooted in the literature surrounding NSF BI assessment (Cotos, 2019; Kamenetzky, 2012; Nagy, 2013; NSF, 2024; Skrip, 2015) and what had been learned through trainings related to the earlier BI resources (guiding principles, BI wizard). For example, researchers often failed to recognize in their BI plans that community partners may expect mutually beneficial outcomes and allow the time needed for trust-building (Bayley, 2023; Henrick et al., 2017) and even "mutualistic relationships" (Coburn & Penuel, 2016). The BI rubric (ARIS, 2023) explains that the researcher should identify the roles and needs in the partnership, provide strong evidence of mutual understanding of these roles, and provide evidence of equitable and fair planning (ARIS, 2023). Ultimately, the aim was to develop a tool of high utility that both BI professionals and researchers could use to strengthen BI plans. It was also thought that such a tool might be useful to researchers engaged as NSF review panelists in assessing proposed BI plans.

Methods/Procedures

The development, refinement, and testing of the BI rubric relied on both quantitative and qualitative methods using three primary approaches:

1. Assess quantitatively the content validity of the rubric through the use of expert panels.
2. Measure the reliability of the rubric quantitatively, using defined BI plans through sets of raters who are representative of the groups of individuals who might be interested in using the rubric.
3. Evaluate the utility, relevance, and value of the rubric through focus groups and open-ended survey responses from researchers.

The study was ruled exempt by the Carleton College Institutional Review Board. The study was carried out in phases from 2020 to 2024 (see Table 1).

Content Validity

In the fall of 2020, the initial rubric was tested by the full ARIS leadership team against two BI plans (see Table 2). Differences in interpretation of ratings and feedback from the team were used to refine the language of the rubric criteria. Following

Table 1. Timeline of Approaches Used in the BI Rubric Study

Timeline	Methods
2019	Initial rubric developed. 2020 interrater reliability (IRR) assessed by ARIS leaders.
2021	Content validity ratios (CVR) computed using a panel of experts.
2021	Rubric refined by ARIS leaders and assessed by a panel of BI professionals; IRR computed.
2024	Rubric assessed by panel of researchers, IRR computed, CVR computed for new criteria or where there had been less agreement.
2024	Focus groups and surveys about rubric involving a panel of researchers conducted.

these revisions, ARIS leaders sought to understand whether other likely users of such a rubric would interpret the criteria consistently and find the criteria relevant. Assessing the content validity of the rubric—that is, the degree to which rubric criteria were "relevant to and representative of the targeted construct for a particular assessment purpose" (Haynes et al., 1995) beyond the ARIS leaders—was critical in the next steps of its deployment in the BI community.

Subsequently, in January 2021, the project sought to establish the content validity of the rubric through a survey administered to a panel of 10 experts who evaluated the degree to which each criterion in the rubric was "relevant" to its intended use (e.g., Zamanzadeh et al., 2015). Criteria deemed "relevant" by a critical number of experts are typically included in a final rubric, while criteria failing to meet this critical level might be revised or discarded (Ayre & Scally, 2014). ARIS leaders recommended 19 individuals who had long-standing expertise in developing and implementing NSF BI plans. Of these 19, 10 agreed to serve on the expert panel and complete the survey. The panel of experts represented a range of roles comparable to who might use the rubric, with 21% of the sample identifying as administrators in higher education, 43% identifying as researchers, and 36% identifying as BI professionals.

The survey was administered to the experts online using Qualtrics with a Google document link to the full rubric to use as reference. The survey asked experts to rate each of the rubric criteria on a 4-point scale ranging from 1 (*not at all relevant in evaluating a BI plan*) to 4 (*highly relevant in evaluating a BI plan*). The rubric criteria were organized into five areas, and for each area, respondents could also provide additional qualitative feedback on that portion of the rubric (e.g., Are the criteria clearly worded? Are the descriptors for each criterion clearly worded? Are the descriptors for each rating level appropriate?).

Analysis, Revisions, and Validity Testing

Following the procedures identified by Lawshe (1975), a content validity ratio (CVR) was calculated for each of the 13 rubric criteria from the 2021 expert panel. CVR uses the following formula, where CVR is the content validity ratio, N is the number of experts, and N_e is the number of experts indicating that a particular criterion is "relevant" by providing a rating of 3 (*quite relevant*) or 4 (*highly relevant*) on a 4-point scale (Figure 1). Values for the CVR can range from –1 (perfect disagreement) to +1 (perfect agreement). CVR values equal to or greater than zero indicate that at least half of the panel experts agreed that a rubric criterion was relevant.

Figure 1. Content Validity Ratio Formula

$$CVR = \frac{(N_e - N/2)}{N/2}$$

Following the initial analysis of both content validity and interrater reliability, the rubric was further refined to address areas where there were differences in interpretation. Specifically, the Scalability, Sustainability and Replicability criterion was removed, and four other criteria were added (Target Audience Alignment, Research-Based, Innovation, and Checklist).

The revised version was tested by two different representative panels. First, a panel of BI professionals (see Table 2) who were ARIS leaders or part of the Organizational Research Impact Capacity (ORIC)[2] cohort program employed it to rate four BI plans. Subsequently, a panel of 20 researchers was asked to score four BI plans and then participate in a focus group that included survey questions related to content validity (n = 16) or (if not available for one of two focus groups) to respond to a survey with the content validity questions (n = 4). Content validity ratios were computed for items where there was less agreement or where previous content validity ratios were low (see Table 2).

Reliability

The reliability of the rubric was tested along with the content validity. For a rubric to be reliable, it is important that different people rating the same documents can arrive at similar results, demonstrating that the scores are reproducible. The rubric reliability was tested at three points, first in 2020 by ARIS leadership team members, again in 2021 with BI professionals participating in the ORIC program along with ARIS leaders, and finally in 2024 by researchers. The ratings were entered into SPSS software to statistically

[2] ORIC is an ARIS program that aims to increase institutional capacity for BI through workshops and support for BI professionals.

measure the reliability, or reproducibility, of the results using the intraclass correlation coefficient (ICC)[3] interrater reliability (IRR) test (Cicchetti & Nelson, 1993).

The initial rubric was used by four members of the ARIS leadership team to score two BI plans. For each plan, scorers were asked to provide independent ratings on a five-point scale (where 1 = *poor* and 5 = *excellent*) for 12 criteria. Analyses were based on the ratings of the four raters who provided ratings for all of the criteria. In this round, all but two of the criteria had a high level of IRR with excellent ICC scores. One of the authors collaborated with a senior research scientist with extensive BI expertise and past foundation experience on revisions before connecting with ARIS leadership team members to make final determinations surrounding changes. Chief among the concerns addressed was balancing sufficient detail for the user of the rubric, including the range and characteristics of the range, while not being too prescriptive in explicating these details. While Q2 ("To what extent do the proposed activities suggest and explore creative, original, or potentially transformative concepts?") received a high ICC score, indicating that different individuals interpreted it similarly, the item was split into multiple criteria to provide adequate detail promoting the research basis for the project, evidence of the effectiveness of delineated BI practices, and innovation of the BI plan. The 2024 ratings were also analyzed using the random-effects ICC model, with nearly identical results, indicating that the ratings are generalizable to other raters with similar characteristics (Koo & Li, 2016).

Following content validity analysis and revisions, the reliability was next tested with seven ARIS leaders along with a convenience sample of four BI professionals involved in the ARIS ORIC program. This round of testing included 12 refined criteria against four sample BI plans. Adjustments were made to the low-reliability items for clarity based on scorer feedback. Four additional criteria were added based on scorer feedback.

The updated rubric underwent a final round of reliability testing with a group of researchers recommended by the ARIS leadership team, with 20 researchers out of 87 participating. Four BI plans were reviewed against sixteen rubric criteria by all 20 researchers, with all participating in either a follow-up focus group (n = 16) or survey (n = 4).

Qualitative Study

In order to understand the rubric's potential utility and relevance for researchers, the panel of 20 researchers was invited to participate in one of two virtual focus groups (n = 16). The hour-long focus group protocol gauged researchers' overall impressions of rubric use, assessed any challenges researchers experienced in using the rubric, and specifically probed areas where there had been less agreement in interpretation of criteria. The first two authors each did an initial coding analysis of the transcribed audio using an emergent coding scheme developed through the constant comparative method (Corbin & Strauss, 1990). The initial coding themes were compared between the coders and triangulated with the quantitative measures, noting where quantitative criteria met or failed thresholds, to arrive at the findings. An online survey was administered to the four participants who could not attend a focus group to gather their interpretations of specific criteria and general impressions. The open-ended survey responses were analyzed thematically with the transcripts using the same coding scheme. Quantitative questions pertaining to content validity were included in both the focus groups and the survey.

Findings

After undergoing testing by BI experts and ARIS leaders, the BI rubric was modified to address areas with lower reliability and content validity, as described above. Ultimately, the resulting 2024 BI rubric shows strong content validity and reliability across the majority of the criteria (Table 2). Recommendations provided by Wilson et al. (2012) and Baghestani et al. (2017), using exact binomial probabilities, guided the determination of the number of experts required to agree to confirm content validity. Content validity was confirmed for 12 of the 16 criteria. The 2021 panel of experts established content validity for 10 of the items. The content validity for the Target Audience Alignment and Evaluation items were established by the 2024 panel of experts. Recommendations provided by Koo and Yi (2016) guided the interpretation of the ICC value, with values less than 0.5 indicating poor reliability, values between 0.5 and 0.75 indicating moderate reliability, values between 0.75 and 0.99 indicating good reliability, and values greater than 0.99 indicating excellent reliability. IRR was found to be moderate to excellent for 13 of 16 criteria.

[3] ICC estimates and their 95% confident intervals were calculated using SPSS statistical package version 28.0.0 based on mean rating, two-way mixed-effects model with absolute agreement.

Table 2. Content Validity (CVR) and Reliability (ICC) of the BI Rubric

	2020	2021	2021	2024	
	ARIS leaders (*n* = 4)	**Panel of experts (*n* = 10)**	**BI professionals (*n* = 4)** **ARIS leaders (*n* = 7)**	**Researchers (*n* = 20)**	
	2 BI plans	***4 BI plans***	***4 BI plans***	***4 BI plans***	
Rubric criterion	ICC	CVR	ICC	ICC	CVR
Q1. What is the potential for the proposed activity to benefit society and contribute to achievement of specific desired societal outcomes?					
1.1 Target audience characteristics	.81	.80	.81	.87	-
1.2 Target audience engagement	.83	1.00	.77	.78	-
1.3 Target audience alignment	-	-	-	.42	.79
Q2. To what extent do the proposed activities suggest and explore creative, original, or potentially transformative concepts?					
2.1a Potential to be transformative	.95	-	-		
2.2 a Scalability, sustainability, replicability	-	.25	-		
2.1b Research based		-		.77	.50
2.2b Evidence based	-	.25	.86	.70	.00
2.3 Innovation		-		.71	-
Q3: Is the plan for carrying out the proposed activities well reasoned, well organized, and based on a sound rationale? Does the plan incorporate a mechanism to evaluate success?					
3.1 Project objectives	.97	.60	.91	.84	-
3.2 Links to NSF target outcomes	.92	.80	.50	.57	.70
3.3 Evaluation	.99	.40	.79	.92	.80
Q4: How well qualified is the individual team or institution to conduct the proposed activity?					
4.1 BI team	.95	.80	.93	.89	-
4.2 Partnership	.89	.60	.91	.93	-
4.3 Partnership needs	.81	.80	.85	.93	-
4.4 Timeline	.94	.80	.92	.97	-
4.5 Checklist	-	-	-	.84	-
Q5: Are there adequate resources available to the PI (either at home institution or through collaborations) to carry out the proposed activities? Is the budget allocated for the activities sufficient to successfully implement them?					
5.1 Infrastructure	.99	1.00	.93	.90	-
5.2 Budget and budget justification	.95	1.00	.35	.65	.40

These findings suggest that BI professionals and researchers consistently rated and similarly interpreted most of the rubric criteria when employing it to evaluate a BI plan.

Different patterns for validity and reliability emerged across the criteria. For example, it is worth noting that the criterion evaluating the qualifications of a team or institution in conducting a proposed BI activity demonstrated excellent reliability and good validity, with similar findings for the criterion evaluating the identification of appropriate infrastructure for supporting BI activities. Whereas the different versions of the criterion evaluating whether BI activities explore creative, original, or transformative concepts demonstrated good

reliability, they were consistently problematic in confirming content validity. Conversely, one of the newer subcriteria, Target Audience Alignment, showed good validity but less reliability with the 2024 panel of experts.

Audience Alignment

Researchers characterized who was considered the "audience" differently when considering the audience alignment criterion in relation to a BI plan. For example, one researcher wondered about the distinction between participants and target audience, noting that "sometimes my participants are my target audience and sometimes they're not." Other researchers wondered whether a "beneficiary" would be the same as an audience and similarly questioned how "stakeholder" relates to the audience. These distinctions were important in knowing whose community needs and interests were described. When the BI activity was related to engagement within a public school setting, researchers said they would use the alignment criterion to assess whether activities showed knowledge of alignment to state or local curriculum expectations. One researcher summarized the challenges of using this subcriterion in this way:

> If I know a lot about that audience, and the proposed plan looks like it would work for that audience, I may rate that highly. Even if they didn't outline how it works for that audience. Whereas, if it's an audience I'm not as familiar with and they didn't support it, and it just smelled funny to me, I may rate that alignment question low, not because I went line by line and looked at their justification of the mechanisms, but just because my personal bias would really play. And I think this question especially, I think, could be subject to bias, at least for me.

Research, Evidence-Based, and Innovation Criteria

Across the different derivations of the rubric, the subcriteria in Section 2 related to how the BI activities explore creative, original, or potentially transformative concepts remained problematic. In both focus groups, researchers engaged in a rich discussion surrounding their thinking as to what constitutes "research-based" versus "evidence-based." Some noted how disciplinary differences might lead to distinctions, and others found it confusing to isolate the two criteria from each other. Still others noted a hierarchy, with one subcriterion being more important in comparison to the other.

Researchers across both groups grappled with the distinctions between the criteria, with one researcher asking, "Isn't research using evidence?" In one group, the researchers noted their opposing points of view, with one stating in relation to the innovation criterion, "I interpreted the evidence-base as being in opposition to innovation. Like saying, you have an existing relationship and you're going to keep doing what already works." Another responded,

> I saw the research base being maybe more oppositional to innovation, … are you only going to participate in BI activities that have a big track record of people doing them already, that have a publication out that you can cite? Or are you going to be trying things that are new?

Another researcher stated that being able to cite prior research and literature that supports a BI approach demonstrates "the criterion of plausibility" that they would constitute as "evidence based."

Disciplinary differences were noted for how to disentangle the two criteria. One researcher stated,

> I do think they are different for social science type of research where you engage the community. And if you use, for example, the community participatory action research approach, you have to know what works, does not work in a local community. You can't just bring in research and say, "We're just going to educate you guys on things."

Another researcher who came from a natural science discipline stated a desire to have high impact and also questioned how much bandwidth they would have to be innovative with their BI plans or even know what counts as novelty.

Other Cross-Cutting Themes

Two other themes related to budget justification and NSF target outcomes emerged across the focus groups and aligned with comments from the 2021 panel of experts. For the budget justification criterion, researchers' feedback related primarily to the provided BI

plans' lack of sufficient budget information rather than the criterion itself. Because the researchers for the study did not have access to full project descriptions and associated budget documents that one would typically have when preparing a proposal or serving on an NSF panel, they stated it was more difficult to evaluate on this criterion. Circumstances would clearly be different in using the rubric outside of the study, where all information would be available. In relation to the NSF target outcomes, researchers across both focus groups unanimously stated that all of the nine NSF outcomes should be explicitly listed in relation to the criterion. (The rubric they reviewed listed three with an asterisk denoting others.) As one researcher noted, "Do we penalize somebody who has focused on improving national security or development of a diverse globally competitive STEM workforce? … This could bias a panel." Others went on to question how critical this criterion was on the rubric because the list of potential NSF outcomes as stated by NSF is not limited to the nine listed in the NSF proposal and Award Policies & Procedures Guide (PAPPG).

Overall Impressions

Researchers across both focus groups and those who submitted reflections in survey format noted a number of overall impressions about the utility of the BI rubric. One major theme that emerged was the usefulness of the tool in various contexts—for example, as a self-guide to conceptualize or judge the completeness of their own BI plans or as a tool that panel reviewers could use as a way to focus and provide consistent feedback. These uses were characterized by one focus group participant as follows, to which many other participants nodded in agreement:

> I found it useful for thinking through it in both directions. When I have been on NSF panels, I felt like no one really knew what was supposed to be in a broader impact statement. And so, it was very hard to judge and evaluate. And I think at the same time when you're writing it, being able to see what's expected as part of the broader impact statement, would make it very. … You can write with much more purpose if you know what is supposed to be there. So, I thought it was useful from both points of view.

Discussion

Criteria With Less Agreement: Challenges of Explicating "Proven" Versus "Innovative"

The preceding section described how the criteria used to evaluate "to what extent the proposed activities suggest and explore creative, original, or potentially transformative concepts," despite significant revisions, failed to meet a standard for content validity. Focus group findings point to differences in interpretation of the three subcriteria based on reported disciplinary background or experience working in specific community engagement contexts. The findings also point to the role a potential bias toward established practices over inventive strategies could play in evaluating BI plans.

Researchers in the 2024 focus groups described different interpretations for the "evidence-based" versus "research-based" and "innovation" criteria. They attributed the distinctions between their views to their varied disciplinary backgrounds. Social science researchers stated the importance of understanding the local context of the BI engagement and argued that this demonstrated understanding is part of the evidence base, whereas natural science researchers reported their biases toward the types of community contexts they had previously worked. The rubric descriptions for each criterion underscored researchers' confusion. They reported trouble disentangling the innovation criterion that needed to align with partner needs, research-based criterion that emphasized prior experience and scholarly literature, and evidence-based criterion that focused on established practices. Researchers questioned what an exemplar BI plan would look like that would encompass and meet high standards along all three subcriteria.

The development of the BI rubric arose from a need for greater clarity surrounding what constitutes strong BI plans. It follows that if less is understood about robust BI engagement, more researchers may favor what is established versus what is truly novel. Other scholars have already highlighted the challenges of funding mechanisms and systems that favor established research areas over those considered more novel (Heinze, 2008). Scholars point to review processes of funding agencies that perpetuate this bias and lead to researchers being more averse to risk-taking (Franzoni & Stephan, 2023; Lane et al., 2022; Lee, 2022). The NSF panel process of weighing multiple merit criteria into a holistic recommendation for funding often involves peer discussion. Such

discussion can propagate negativity bias, where panelists lower their own scores to norm with other panelists, thus favoring more conservative proposals. More vocal panelists may elevate their own particular idiosyncrasies or experiences, thus favoring particular established methods (Lee, 2022; Zoller et al., 2014). The existing criteria for this aspect of the rubric balance innovation with partner needs and leverage documented practices and existing scholarly literature. Yet the focus group discussion for these elements draws attention to the challenges of weighing individual criteria and the different interpretations and bias they could introduce.

Conclusions: Features for Use

The BI rubric development and testing involved key informants representing BI professionals with experience supporting pre- and postaward proposals and researchers across many disciplinary fields. All informants agreed on the efficacy of the rubric's potential use. Understanding of the range of uses was strengthened through the involvement of both BI professionals and researchers as part of instrument development. The feedback from this study highlights the strengths of the rubric, potential challenges to consider, and the utility of the rubric in different contexts. The features for use reported by the informants include the following.

- The rubric presents a common set of criteria that codify expectations for BI. Common expectations would be useful to those serving on NSF panels in giving strong dimensions on which to evaluate proposals, hold meaningful panel discussions surrounding BI, and strengthen feedback to proposers. No instrument alone can mitigate the bias inherent in a review process. The instrument presents a starting point for setting more common expectations for stronger BI.
- Researchers grappled with how one would weigh the set of criteria to arrive at a single score. It is unclear whether providing guidance for such use would exacerbate or minimize the commensurate bias of the summative score process (Lee, 2022). Arriving at a common group judgment for a proposal implies a high-stakes assessment. Caution should be exercised in any consideration of using the rubric in a cut-score implementation where validity was not tested for these circumstances.
- Nearly all of the criteria met validity and reliability standards. The aspects where reliability or validity was not as strong (e.g., audience alignment and Section 2) point to differences in language usage and interpretation. The discussion among researchers related to the audience alignment criterion centered on different interpretations for whose needs and interests are met. Examples of community-engaged scholarship exist along a continuum (e.g., service learning, participatory action research, community-based learning, public information networks, community–campus partnerships) where societal benefits may include community partners and the institutions involved (Fitzgerald et al., 2017). So it is not surprising that researchers may have different interpretations as to whose needs and interests are met. Differences in interpretations for Section 2 highlight both distinctions in disciplinary fields interpretation and inherent biases in peer review processes that may favor more conservative approaches (Lane et al., 2022). Revising Section 2 to meet content validity across all the epistemologies of potential users may not be feasible. Yet the imperfect criteria demonstrate the value of engaging researchers in thinking deeply about evidence, scholarly literature, and what constitutes novel within a BI plan.
- Finally, all the key informants reported on the range of uses and flexibility of the rubric. The organization of the rubric into five common questions with associated criteria was credited in strengthening its flexibility for varied use. BI professionals viewed it as an effective professional development tool that they could use with researchers on their campus. They envisioned being able to delve into one or more criteria in trainings as needed. Researchers saw themselves using it to evaluate and develop their own BI plans within a proposal, as a panel reviewer of other proposals, and as a metric for gauging their implementation of a BI plan should the research be awarded.

These possibilities for BI rubric use can serve to strengthen guidance for BI plan development, meaningful discussions and decisions surrounding assessment of BI for funding processes, and implementation of BI plans. We also hope that the

rubric will engage the community in conversations that lead to the clarification and improvement of policy around BI at NSF. Providing greater clarity surrounding BI opens opportunities to bolster partnerships supporting BI, foster improved community engagement scholarship, and ultimately strengthen the benefits to society in significant ways.

References

America Competes Reauthorization Act of 2010. Pub. L. 111-358. (2010).

Advancing Research Impacts in Society. (2020). *BI guiding principles 2.0*. https://researchinsociety.org/resource/guiding-principles-2/

Advancing Research Impacts in Society. (2023). *Broader Impact plan rubric*. https://aris.marine.rutgers.edu/BI%20Rubric%202023-11-16.pdf

Ayre, C., & Scally, A.J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 79–86. https://doi.org/10.1177/0748175613513808

Bayley, J. (2023). *Creating meaningful impact: The essential guide to developing an impact-literate mindset*. Emerald Publishing Limited.

Baghestani, A.R., Ahmadi, F., Tanha, A., & Meshkat, M. (2019). Bayesian critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *52*(1), 69–73. https://doi.org/10.1080/07481756.2017.1308227.

Cicchetti, D.V., & Nelson, L.D. (1993). Re-examining threats to the reliability and validity of putative brain-behavior relationships: New guidelines for assessing the effect of patients lost to follow-up. *Journal of Clinical and Experimental Neuropsychology*, *16*(3), 339–343. https://doi.org/10.1080/01688639408402644

Coburn, C.E., & Penuel, W.R. (2016). Research–practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, *45*(1), 48–54. https://doi.org/10.3102/0013189X16631750

Corbin, J.M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Cotos, E. (2019). Articulating societal benefits in grant proposals: Move analysis of Broader Impacts. *English for Specific Purposes*, *54*, 15–34. https://doi.org/10.1016/j.esp.2018.11.002

Doberneck, D. (2016). Are we there yet? Outreach and engagement in the consortium for institutional cooperation promotion and tenure policies. *Journal of Community Engagement and Scholarship*, *9*(1), 8–18. https://doi.org/10.54656/RNQD4308

Fitzgerald, H.E., Van Egeren, L.A., Bargerstock, B.A., & Zientek, R. (2017). Community engagement scholarship, research universities, and the scholarship of integration. In J. Sachs & L. Clark (Eds.), *Learning through community engagement*. Springer. https://doi.org/10.1007/978-981-10-0999-0_03

Franzoni, C., & Stephan, P. (2023). Uncertainty and risk-taking in science: Meaning, measurement and management in peer review of research proposals. *Research Policy*, *52*(3), Article 104706. https://doi.org/10.1016/j.respol.2022.104706

Haynes, S.N., Richard, D.C.S., & Kubany, E.S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Heinze, T. (2008). How to sponsor ground-breaking research: A comparison of funding schemes. *Science and Public Policy*, *35*(5), 302–318. https://doi.org/10.3152/030234208X317151

Henrick, E.C., Cobb, P., Penuel, W.R., Jackson, K., & Clark, T. (2017). *Assessing research practice partnerships: Five dimensions of effectiveness*. William T. Grant Foundation.

Kamenetzsky, J.R. (2012). Opportunities for impact: Statistical analysis of the National Science Foundation's broader impacts criterion. *Science and Public Policy*, *40*(1), 72–84. https://doi.org/10.1093/scipol/scs059

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lane, J.N., Teplitskiy, M., Gray, G., Hardeep, R., Menietti, M., Guinan, E.C., & Lakhania, K.R. (2022). Conservatism gets funded? A field experiment on the role of negative information in novel project evaluation. *Management Science*, *68*(6), 4478–4495. https://doi.org/10.1287/mnsc.2021.4107

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lee, C.J. (2022). Commensuration bias in peer review. *Philosophy of Science*, *82*(5), 1272–1283. https://doi.org/10.1086/683652

Nagy, D. (2013). *Evaluating the broader impacts of sponsored research through the lens of engaged scholarship*. (Publication no. 3589848). [Doctoral dissertation, University of South Dakota]. ProQuest Dissertations & Theses Global.

National Alliance of Broader Impacts. (2015). *Broader Impacts guiding principles and questions for National Science Foundation proposals*. https://eso.stanford.edu/sites/g/files/sbiybj18016/files/media/file/broader_impact_guiding_principles.pdf

National Science Board. (1997). *National Science Board and National Science Foundation staff task force on merit review: Discussion report*. https://www.nsf.gov/nsb/documents/1997/nsbmr975/nsbmr975.htm

National Science Board. (2011). *National Science Board National Science Foundation's merit review criteria review and revisions*. https://www.nsf.gov/nsb/publications/2011/nsb1211.pdf

National Science Board. (2020). *National Science Board Vision 2030*. https://www.nsf.gov/nsb/publications/2020/nsb202015.pdf

National Science Foundation. (2024). *Proposal & award policies & procedures guide (PAPPG): NSF 24-1: Effective for proposals submitted or due on or after May 20, 2024*. https://new.nsf.gov/policies/pappg/24-1

Skrip, M.M. (2015). Crafting and evaluating Broader Impact activities: a theory-based guide for scientists. *Frontiers in Ecology and the Environment, 13*(5), 273–279. https://doi.org/10.1890/140209

Wilson, F.R., Pan, W., & Schumsky, D.A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197–210. https://doi.org/10.1177/0748175612440286

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.-R. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, *4*(2), 165–178. https://doi.org/10.15171/jcs.2015.017

Zoller, F., Zimmerling, E., & Boutellier, R. (2014). Assessing the impact of the funding environment on researchers' risk aversion: The use of citation statistics. *Higher Education*, *68*, 333–345. https://doi.org/10.1007/s10734-014-9714-4

About the Authors

Ellen R. Iverson is the director of the Science Education Resource Center at Carleton College. Kristin O'Connell is an evaluation/education associate at the Science Education Resource Center at Carleton College. Janice D. McDonnell is an associate professor at the Rutgers University School of Environmental and Biological Sciences and the science engineering and technology agent for the department of 4-H Youth Development. Susan D. Renoe is associate vice chancellor for research and assistant professor of strategic communication at the University of Missouri and executive director of the Center for Advancing Research Impact in Society. Liesl Hotaling is the president of Eidos Education.