Option 1

David MacD

I'm concerned with a few aspects of this approach.
(1) My biggest concern is "user specific context". under adaptive requirements. I think its really hard to get inside the end user's experience. Context is complicated, has endless variations, is individual, and is difficult to measure. It puts too much power in the hands of the evaluator. It's like giving the evaluator the ability to move the goalposts, and the author has almost no recourse. This is particularly worrisome if the evaluator is on the opposite side of a lawsuit as the company. How can a company defend itself in this case? There is no objective measure to point to.
(2) Protocol based testing: I think it's very difficult to get inside a company to determine if they followed a particular protocol.

I think both of these could be included as advice.

Gregg

I am unable to use the ratings above. Many of these are "sort of" or "sometimes but not others" etc.
Everything rated above as *YES - BUT PREFER NOT* = means -- it is sometimes true and sometimes not
So I am mostly going to leave comments.

Overall - *LOTs of good ideas and concepts* - too many to list here
I used this OPTION to build my Option 2 - addressing the concerns I has with this and adopting all the good ideas from

My concerns - that I addressed in Option 2
[These are concerns with several of the options below. I will list them here and refer to them below]

1) *USE OF VIEWS* - -which I was a proponent of but we could never make them work. So I suggest we explore this up front and see if someone has found a solution to the problems - and if yes - go for it but if no -- drop it (until someone does) since it is distractingly attractive but diverts us from solving problems with next best approach (page, doc, software) as was used in WCAG2ICT.

2) *PARTIAL or % or SCORE IS GOOD ENOUGH* - Use of "essential" or "most" as being ok for conformance. This goes along with scoring. I am happy with scoring etc. for extra credit but not for basic conformance. Too easy to game and have "95% but not the front door" situations. And to leave some groups out.

3) *USER NEED based* Anything *user-specific* has the problem that - before I release my product - I need to know exactly which users' I am supposed to meet the needs of. And how to turn that "need" into a specification I can meet. There are dozens of different types of "blind" people alone - who have different needs and can use or not use different things. Tell me what my product must do or not due to meet the needs you are thinking of -- don't give me a vague description of a need unless it is "users need your product to do (or not do) this". In which case, don't call it a user need. Call it a specification. And don't tell me how to do it - just tell me what the result should be -- what the outcome should be. (Hence use of Outcome measures is great. and use of 'user's needs is a problem)

4) CONTEXT - This talks about adaptive tests that are dependent on "context". This is a variation on user needs. If things are context based you need to tell me - in advance of my releasing my product and someone evaluating it - exactly what the contexts are -- and what you think the changes in the outcome measures are for each context. Once you have defined several contexts - don't call them contexts but 'conditions' and make the provision conditional. " If this condition (context) exists then your product/site must do this. For this other condition (context) it must do that.

5) EXTENSIBILITY - This talks about "when multiple specifications or standards are available for the same test". If they all come up with the same result - but just use different measurement spaces that can be translated / converted into each other -- then just name one - and the others can all be converted into it. If different ones are appropriate (and the others are not) in different situations (say color spaces) then make it conditional (if in this colorspace use this - if in that use that) OR (use the test appropriate for the colorspace. NOTE here is the test for this colorspace and that.). But there should never be two test that both are acceptable and give different results. If that is the case then you must say it must pass both tests. In all cases there is no EXTENDING of any requirement.

6) QUALITATIVE TESTING used for CONFORMANCE -- I think qualitative testing has a role -- for going beyond basic conformance (Adjectival testing and recognition of progress toward or progress beyond conformance (pass). But qualitative can't be used for basic conformance and still maintain the objectivity and inter-rater reliability that is needed for the requirements to be used in regulatory work when we are done

7) USER TESTING - Again - this is great and should be done on all sites. But the result is completely dependent on the users used to test. Unless the exact users are specified and then never change (don't change users and users don't themselves ever change) you can't test your product before release and know what they results will be after. So user testings should definitely be in our recommendations and we should include a way to recognize and reward user testing - but it can't be part of basic conformance.

BruceB

My understanding is that to get to Silver, one or more "extensible requirements" must be met. It may be the case that W3CAG ends up with a bunch of these, in which case only needed one might be too low of a bar.

Rachael

I think it is realistic from a user's perspective but may be hard from a guidelines perspective (AKA this would move the heavy lifting to the working group). As written, it does not motivate people to move between levels.

Makoto

At this moment, I'm not sure if it is realistic or not. For example, the example of "Guideline: Provide text alternatives for images" includes the following items:
- industry or organization standards
- assistive technology testing
- user testing

It is still ambiguous what "industry or organization standards" are and how we must conduct "assistive technology testing" or "user testing". And it will needs more discussions on the definitions of these items.

For example, I came up with several questions such as:
- How many testers must we have at least to do the "assistive technology testing" or the "user testing"
- The testing results depend on testers" skill, literacy and experiences for IT, Web, AT etc. Is it okay even if the results are not reproducible?
- How can we measure the good enough quality of the "assistive technology testing" or the "user testing" to determine if it achieves Gold?
- Will we provide the guidance on how to conduct these kinds of testing at last?

If the answers for these questions are already available, my answers for this survey might change.

And I might not have catched up with all discussions on this issue which have been already done. Sorry about that.

Rain

In my experience before Google, when I was working with consulting and web dev agencies, I repeatedly encountered the goal to be "let's meet the bare minimum required to not get sued." I feel that this model is able to be used by agencies to do a lot less, not more. I'm also struggling with why something can achieve a "silver" or "gold" rating when only meeting 1

test at that level. The content producers could choose any one they want, at the exclusion of anyone who needs other supports.

Jeanne

We don't know the metrics (validity, reliability, complexity) yet until we test it. Overall, Option 1 has some very good ideas but it is still at the complex phase. I think it can be simplified to make it easier to understand. I like that Bronze, Silver, and Gold are different categories, although I worry about using test types to stratify the levels. I think that it could have problems from a civil rights experts, even though it makes perfect sense to a group of developers.+

Wilco

The conformance model relies on the tests described in the methods. The outcome alone does not tell things like which tests are at bronze, and which are silver, or how to decide if something's a single issue, or multiple. If the only way to know if a site is conformant is by using AGWG's methods, then those methods will need to be normative. That feels like it could be a problem.

I'm not keen on the "may have no more than 3 errors" idea. Firstly, I don't think it's on the W3C to decide how many accessibility errors in an acceptable number of errors. Doing this also creates a situation where having 3 problems is just as good as having no problems, which doesn't feel problematic.

It also raises an enormous challenge in us now needing a consistent way to know what is one error, and what is two. For example, if two paragraphs have bad contrast, is that one error, two, an error for every word, every letter? Who is going to decide these things? One of the lessons learned from ACT was not to try and tackle this. There be dragons.

MichaelG

The problem with the responses you provide (basically "yes, no") is that they're pretty binary when I don't think my assessment leads me to that. Concepts like 'unsubstantiated' or 'unclear' would have been useful to really allow me to benefit more from this exercise.

Option 2

Gregg

This is my proposal - so all my comments are embedded in the OPTION description - including these items above.

BruceB

Gold as needing only 75% Good (or better) seems low, but the tiers can be adjusted as we go.

Rachael

My main concern is that the combination of percentage and adjectival will hide accessibility issues. I am also concerned that the coga standards will end up as assertions and we will mimic the equity issues we have in 2.x.

Rain

The adjectival ratings introduce subjectivity, which is both a positive (supports greater equity), and also a challenge (requires more knowledge of testers, potential risk to inter-rater reliability). Overall, I feel that this model has potential to support more nuance and change in an ever-shifting technology landscape.

Jeanne

There are some very good ideas in this proposal that solve some problems, so it should move forward, but it gets very tangled at the test level. Like Option 1, it also needs some simplification before it can progress further. It may be that different terms are being used for the same thing, but expectations, requirements, outcomes seem to be used interchangeably for the same concept. This needs to be teased apart. The Scoring section has Adjectival Reporting that is a very good idea. Adjectival by category has some serious flaws, and would need testing. The Bronze Silver Gold proposal probably would be approved by civil rights experts, but has questionable metrics and would require prototyping and testing. There is so much emphasis on making different types of test pass/fail, that it misses the value in using the tests. So, overall, it has some good ideas, but needs a lot of work.

Wilco

((skipped, proposal is 18 pages long))

Option 3

Gregg

NOTE: *YES - BUT PREFER NOT* in ratings above means -- it is sometimes true and sometimes not

VERY SIMILAR to Options 1 -- so see those comments

Picked up the concept from Option 2 regarding Adjectival scores above and below conformance. However this version has only 1 level below conformance. So it is not able to measure process toward conformance. if you fall short - you can't report getting closer to conformance without moving from that level all the way to conformance. I recommend 2 below and 2 above rather than 1 below and 2 above.

Bruce

I feel like Alastair's poor is my unacceptable, and Alastair's okay is my good. Otherwise, general concept seems sound.

Rachael

My primary concern with this approach is consistent results across testers. Defining the bands would be very difficult.

Rain

I'm concerned that this model is too descriptive and subjective and would be interpreted differently by everyone applying it.

Jeanne

I'm guessing about the validity (since we haven't tested it yet) but I suspect that choosing the lowest test result for the outcome level will not be valid. But that is a minor point and we could work more on that. I really like the scoring idea -- it's much easier to understand. The Bronze Silver Gold proposal probably would be approved by civil rights experts.

Wilco

It's not clear to me how someone could score an outcome consistently without those methods. It seems to me methods describing the different levels are required, and so must be normative.

It also seems to me this is going to get us into "image counting" territory again. I feel like if we wanted to have this level of granularity in WCAG, then we might as well go back to success criteria in different levels like we have in WCAG 3.

Option 4

Gregg

NOTE: *YES - BUT PREFER NOT* in ratings above means -- it is sometimes true and sometimes not

Lots to like here
Builds on and improves Option 1 - incorporating Option 2 ideas - (and adds new ideas I would like to add to Option2 if there is a other round)
Some HIGHLIGHTS
1) This focuses on OUTCOMES -- good
2) it restricts SCORING to the quality and methods (adjectival etc.) and does not use it for conformance
3) Introduces Badges to provide progress between levels of Bronze Silver Gold
4) Bronze focuses on Conformance - Silver and Gold on quality and doing more.

PROBLEMS/ CONCERNS
- This option preserves VIEWS -- see comments above. Good if we can figure out how to make VIEWS work but we have not been able to.

- it includes ability to "pass" if you get X of Y - this raises the PARTIAL IS GOOD ENOUGH problem discussed above. "What if whole site is accessible except for 5 % but that 5% are showstoppers for one or more disabilities.?" or what if the shopping site is accessible except for all the sale coupons or advertisements of better products.

- it includes ADAPTIVE and EXTENSIBLE tests. These have the problems cited above

- it talks about choosing different methods for conformance. This looks like techniques and should not be in the standard -- since they need to be updated regularly to account for new technologies - and standards are frozen. Suggest keeping techniques out of the standard. And there should not be multiple ways to test. Conditional ways maybe but not multiple unless ALL must be met. (else no rater agreement since they can choose different measures).

QUESTIONS
I do not understand what "Aggregate outcomes" means as used here - nor do I understand the question posed:
" aggregate Outcomes, where the methods define the scoring for each outcome.
Open Question that needs exploration: Issues that interplay
- Can all thresholds be set lower to account for this?
Or, can they be kept within a single Outcome? "

Bruce

This approach seems the most distinct from from the other five.

Rachael

My main concern is that the coga standards will end up as assertions and we will mimic the equity issues we have in 2.x.

Makoto

It looks a little bit complicated. But it is worth continuing further discussion. I'd like to see how it works for other outcomes (alt text, heading markup, color contrast, keyboard, etc.)

Rain

Reading this reminded me that it would be helpful if we could shift our language. Instead of "users who can't hear," could we say "users who require visuals for understanding" -- framing the language based on what people can do instead of what they cannot do. This is a general statement about W3C language, not specific to this option.

Jeanne

This is an innovative approach with a lot of good ideas. Badges to encourage moving up to Silver level is more motivational, IMO. 5.1 Scoring tests idea was strongly discouraged in the responses to the FPWD. Responders don't want scoring to vary by method, that makes it too complex to remember. The Bronze Silver Gold proposal probably would be approved by civil rights experts.

Wilco

I had a difficult time understanding this one. I'm not really clear on what badges actually are in this proposal. What I'm reading in here that i'd love to explore further is the idea of having "opt-in" requirements. Things not required by default, but if you claim your site meets them, then they become required for a test. It could be a great way to handle new technologies, and maybe deal with subjects like accessibility for kids.

I think the risk of it though is that organisations may not feel motivated to pick up more than the base level. I do worry that the badges could make things pretty complex, and having to work out which are the right methods to test could result in a lot of overhead for tests.

Option 5

Gregg

NOTE: *YES - BUT PREFER NOT* in ratings above means -- it is sometimes true and sometimes not

HIGHTLIGHTS
- required items are Bronze -- silver and gold add % of optimal methods [same as options 2, 3 and 4]
- mixes REQUIRED with OPTIMIZED so OPTIMIZED are seen

CONCERN
- talks about *Pass Conditional* mixed in with REQUIRED but not clear what that means or if it is objective or would have high inter-rater reliability.

QUESTIONS
- this is more of a concept than a full option. Suggest combining concepts with other option = or adapting it to enhance other option

Bruce

Option 5 does not seems as fleshed out as the others (just slide notes, not a separate doc from template).

From the captioning example provided, I really like how thoughtfully Bronze/Silver/Gold is awarded depending on the Methods applicable to the Guideline. This approach is sound, but seems like an additional level of complexity.

Rachael

I am not I fully understand "pass Conditional with Rating" but I think this approach has a lot of potentials.

Rain

Not clear on how the scoring works here so am not confident in my responses, so chose "no" for scoring and conformance.

Jeanne

This approach has a lot of potential and needs more details. I like the categorization of "Required" and "Optimized". That is a solid basis for organizing and is simpler than adjectival. We have to be very careful balance the needs of disability groups in Required vs Optimal. This has to be based on user needs otherwise we are perpetuating structural problems in WCAG2. The Bronze, Silver, Gold isn't the best of the options, and has potential

for not being accepted by civil rights experts because it is test oriented. I think it could be improved, however. This needs more work, but should go forward.

Ranking Question

Laura

Option 2 is the most realistic to me. All of the others seem too complex.

However, a drawback to option 2 is that it continues the same type of A/AA (minimum requirements) or AAA (going beyond) as WCAG 2.X . A problem with that is that many won't go beyond what is required. If option 2 could be combined with something such as John Foliot's 2021 proposal where you have to do everything and you are scored on how well you do it it might help.

Side note: Evaluating user processes and flows will be a problem for all of the proposals. It would have a huge impact on evaluation testing at scale.

Gregg

Best I think is OPTION 2 (which drew heavily from OPTION 1)
- with Badges concept from OPTION 4
- and some of OPTION 3 which reduced bel
- and looking at OPTION 5's structure of mixing Required and Optimal (and recommended?) into a GROUP (general guideline?)
- and exploring the "criteria of evidence to support the assertion" from OPTION 6

Bruce

I ranked 4 highest because it seems like the model that best provides motivation for (1) sites that are just getting started with accessibility (but not close to AA) but also (2) sites that are AA and looking to do even better.

Given that several of the proposal use adjectives, I think we should discuss tiers. My thinking is that there is: unacceptable, poor/weak/tolerable (less than AA/Bronze), okay/marginal (close to AA), good (AA, and nothing marginal), and very good (AA, many AAA).

Suzanne

I think any of these would be fine place to start. I worry that having only one level (bronze) to house true/false requirements will become an issue that will block release and create tension, since there will likely be a very large number of true/false requirements stemming

from a goal to cover as much of the FAST document as possible. (Examples: 3D printable files; tactile printable files; haptics; visual instructions that do not require reading for general audience products) If a requirement that helps a small group has to move from bronze to gold, that will be frustrating. But having to remove it completely or have a standard with too many requirements to be realistically adopted would be much more frustrating. Gold and/or Silver should also be able to house true/false. In addition: I think a lower level than Bronze was stated as a requirement for Japan. And an innovation level to house those things we can't expect at this time would be nice.

Makoto

After I'm done with this survey, I'm feeling that I'd like to see how the WCAG 3 Outcomes/Methods will look like to discuss the conformance model. If we can see how the existing SC in WCAG 2 and additional/new ones will look like in WCAG 3, we might be able to discuss the suitable conformance model more effectively. That's what I'm feeling now.

Plus, they might need a lower level than Bronze to promote making digital contents more accessible in countries where web/digital accessibility has not become legal requirements

Rain

Unclear on Model 5 so not marking as "don't want" or ranking. There may be something to the conformance that is being suggested here if it is more clearly defined.

In general, for *all* of these, I think it would also be helpful to have the conformance rating based more on success with entire tasks or jobs within the product rather than units. Instead of trying to conform to a % of optimal methods, moving more towards, "users can complete all tasks with x-ease" or something more closely tied to success with the actual purpose of the product than with individual bits or components.

Essentially, it would be great to see the conformance and scoring based more on *outcome* and less on *component interactions.*.

Jeanne

Option 1 needs simplification but overall structure is excellent. Option 2 has a very good idea for scoring adjectival and for Bronze, Silver, Gold. Option 3 has a very good approach to organizing the testing. Option 4 badges are more motivating than any other proposal and should go forward. The edits to the intro to make them more designer and developer oriented should also go forward. Option 5 required and optimized is a simpler and easier to understand way of organizing the guidelines and outcomes.