NEW AMERICA

December 2020

Reflecting on the Digital Standard

Andi Wilson Thompson, Nat Meysenburg, & Ross Schulman

Acknowledgments

In addition to Ranking Digital Rights and Consumer Reports, we would like to thank the many product designers, developers, and manufacturers that provided feedback on the Digital Standard. We would also like to thank Craig Newmark Philanthropies for its generous support of this work. The views expressed in this report are those of its authors and do not necessarily represent the views of Craig Newmark Philanthropies, its officers, or its employees.

About the Author(s)

Andi Wilson Thompson is a senior policy analyst at New America's Open Technology Institute where she focuses on issues including digital security, vulnerabilities equities, encryption, and internet freedom.

Nat Meysenburg is a technologist at the Open Technology Institute who works on building and maintaining systems with privacy, security and freedom in mind.

Ross Schulman is a senior counsel and senior policy technologist at New America's Open Technology Institute.

About New America

We are dedicated to renewing the promise of America by continuing the quest to realize our nation's highest ideals, honestly confronting the challenges caused by rapid technological and social change, and seizing the opportunities those changes create.

About Open Technology Institute

OTI works at the intersection of technology and policy to ensure that every community has equitable access to digital technology and its benefits. We promote universal access to communications technologies that are both open and secure, using a multidisciplinary approach that brings together advocates, researchers, organizers, and innovators.

Contents

What is the Digital Standard?

Over the years, product safety testing has been standardized for all sorts of categories of consumer goods. From cars to cribs, increased testing to ensure that products are adhering to best practices in safety has turned many types of products from possibly deadly, to generally safe and trustworthy. For example, the implementation and testing of seatbelts and airbags has dramatically reduced injuries and deaths in auto accidents. These standardizations of best practices often happened in reaction to **observed safety issues**, and an inability of the buyer to evaluate the **comparative safety** of products they are purchasing. The growth in popularity of internet-connected smart devices presents a similar problem. These devices are vulnerable to a new range of security and privacy threats; ones that your refrigerator and slow cooker have never had to face before. Previously, a refrigerator may have been tested for temperature regulation and energy efficiency. Now, it and many kitchen appliances are connected to the internet, allowing you to check what's in your fridge from the grocery aisle, or turn down your slow cooker from a phone app.

Having built reputations based on mechanical, electrical, and design expertise, many manufacturers who have made versions of a product for decades, are now faced with a completely new category of product safety that they've never had to consider. A coffee maker never used to be able to gather personal data, or be made part of a botnet. The risks that these more advanced "internet of things" (IoT) products pose to consumer privacy and security calls for the implementation of a standardized testing framework, so that products can be evaluated and compared.

Such a testing framework is already publicly available through the **Digital Standard**. With the goals of enabling consumer organizations to "test, evaluate, and report on whether new products protect consumer security and privacy," and helping consumers to "make smarter choices about the products they buy," the Digital Standard is working to "create a digital privacy and security standard to help guide the future design of consumer software, digital platforms and services, and Internet-connected products." The standard is composed of 35 tests on issues ranging from encryption, to data retention, to identity policies. Each of these tests includes specific criteria and indicators to review when conducting the test. Most of the tests have existing criteria and indicators listed, however a few of the tests remain “under development” or “under discussion,” which may mean that some aspect of the test language is currently incomplete or nonexistent.

Who Created and Maintains the Digital Standard? Who can Contribute?

The development of the standard was led by **Consumer Reports**, **Disconnect**, **Ranking Digital Rights**, **Aspiration**, and the **Cyber Independent Testing Lab**. Each of these groups applied their experience creating rigorous testing and ranking systems to imagining how to evaluate the huge range of newly emerging IoT devices, which can include everything from door locks, to garden sprinklers, to televisions, to kitchen appliances. The standard also seeks to include expertise from contributors outside of the founding organizations. The open source standard text is **posted on GitHub** and is open for comments, additions, and revisions from any member of the tech community—researchers, developers, advocates, and hobbyists.

The criteria contained in the standard are rooted in the principles that "electronics and software-based products should be secure, consumer information should be kept private, ownership rights of consumers should be maintained, and products should be designed to combat harassment and help protect freedom of expression."

These principles serve to frame testing of digital products around security and privacy protections for the humans using a product. The Digital Standard is designed to make sure that sensitive data is handled properly, rather than a more limited testing framework that may focus only on functionality or durability. By expanding product testing to include the indicators contained in the Digital Standard, consumers can learn not only "will this product actually cook dinner?" or "will this product last?" but also "can this product be easily hacked?" or "will this product leak sensitive data?"

Why is Testing Important?

The huge expansion in the number of connected IoT smart devices over the last decade has made accessing parts of our homes from an app a routine part of daily life. While someone installing a new smart doorbell may not think of it as an internet-connected computer, it is. All of these devices are potentially susceptible to many of the same kinds of vulnerabilities as laptops and phones are. IoT devices also present new possibilities for cyber attacks that can escalate to consequences in the physical world. If your smart lock gets compromised, someone could remotely unlock your front door; if your thermostat gets compromised, your heat could be turned off on a cold winter night. However, the average consumer is likely more accustomed to thinking about the harms associated with the food in their refrigerator spoiling than the computer in their refrigerator getting hacked.

It is often unclear, even to the technically savvy, what pieces of technology were used to make a traditional product "smart," or how closely those pieces integrate into or intersect with the basic functionality of a product. There is not one single way for a manufacturer to go from making thermostats to smart thermostats. Different smart products will connect to the internet in different ways, and present both unique features, as well as unique security risks.

Unlike laptops and phones, a huge number of IoT devices exist in a space where their manufacturers are either unable or unwilling to update a device's vulnerable code. This presents a problem for durable goods, such as major kitchen appliances, which are often expected to function for over a decade, far exceeding the traditional support cycles of most internet connected products. There is often no way to know ahead of time whether a manufacturer can or will provide ongoing updates, or for how long they will do so. There is also little way to know from the outside of the box if a product handles user data correctly.

Many IoT products may not even be tested for their digital security before being sold. And even in cases where manufacturers do rigorously test their products, it is rare for either the testing criteria or the results to be released.

This puts consumers in a situation where they cannot fully evaluate the quality of products, or make informed choices about the risks those products may pose to the privacy or security of themselves or their loved ones, and whether the benefit of the product's features truly outweigh those risks.

An open, shared, and widely-used testing framework could change all of these dynamics. Better testing could allow consumers to make more informed choices, and interacting with those results could help educate consumers on some of the risks inherent with internet-connected things. Widespread product testing could also encourage companies to learn about and implement best practices for digital security.

Why was this Testing Handbook Necessary, and Who is it For?

While security researchers do end up testing a variety of products, their tests are often created to test for specific vulnerabilities found in one product or product line, rather than a comparative evaluation of similar products. The Digital Standard creates a framework for comparative analysis of IoT products by doing what its name suggests—standardizing these types of tests under a single rubric. The goal is that consumer groups, manufacturers, students, and interested hobbyists alike will be able to use the standard to conduct their own testing of products or product types.

However, as it exists, the standard does not provide interested testers all of the information they need to implement the testing protocol. Although many of the tests include procedural overviews which offer helpful guidance as to how a tester might evaluate a specific indicator, they are often broad and incomplete; in many cases leaving it unclear as to what the steps for evaluation may be.

In earlier Open Technology Institute (OTI) projects about the Digital Standard we spoke to many stakeholders in civil society and the private sector who were interested in using the standard for product development or evaluation. The most common response we received was that there was not enough procedural guidance for non-experts to run the tests. Implementing the standard would require testers to possess expertise in a wide range of areas of technology privacy and security policy, and to expand upon the standard's existing procedural overviews. Stakeholders wanted some sort of handbook or guide that would empower them to use the standard as imagined by the developers.

OTI started the **testing handbook** project to provide such a guide. In it we build upon the existing procedural overviews in the Digital Standard by providing detailed step-by-step directions on how to perform the tests. In many cases our work involved expanding the procedures listed in the overview, as well as fleshing out the discussion of more sophisticated best practices or technical approaches that could be used as part of testing. Following some of the testing processes requires familiarity with legal or technical language. Others require testers to choose between various approaches to running the tests or evaluating the results. Nonetheless, our hope is that the step-by-step instructions in the testing handbook will make it easier for other interested testers, including consumer groups, hobbyists, and the product manufacturers themselves, to begin conducting their own product testing.

How does the Handbook Score Products?

One of the key choices that we had to make was developing a rating system for use when reviewing the products. Although a numerical or other type of complex rating process can provide valuable nuance, we decided that it would be too subjective for testers. Ultimately we concluded that products could receive "Pass," "Partial Pass," "Fail," and "Not Applicable" grades based on their performance on each indicator. Not Applicable ratings are reserved for cases where the product does not include the functionality or feature being evaluated in the test. For example, some tests may specify examining a browser component, and many IoT products do not have one; for that product all browser component tests would be marked Not Applicable. Partial Pass ratings are used when an indicator contains multiple requirements and a product does not fulfill all of them. For example, a test on privacy policies could include a list of required best practices, some of which are met by the product's policy, which would result in a Partial Pass grade.

How did we Pick the Products? (And Why aren’t We Naming Them?)

In order to design a handbook that would be valuable to testers, we evaluated a selection of sample products using the Digital Standard. This experience allowed us to write out and then refine the steps required to perform each test, and develop further procedures as necessary. Crucially, we also developed specific criteria for a product to pass or fail the test, so that various testers were using the same method of measuring success.

Ultimately, we tested three products, varied enough to cover a wide range of features offered in common smart devices and selected based on the unique privacy and security threats that the potential failure or compromise of a product could pose. Since not all indicators in the standard fully apply to every product, we also selected a range of products to make sure we covered every indicator at least once. For example the **Known Exploit Resistance test** suggests examining the browser component of connected devices, and while this is important for tablets or even some smart fridges which ship with web browsers, a large number of IoT devices simply don't include a browser feature.

We opted to test products offered by larger and well-known brands, rather than by startups or small manufacturers. We did this in the hope that larger companies may have put more resources into legal documents like terms of service and privacy policies, and that they also may have longer track records with things like patching known vulnerabilities. Since we were not conducting a comparative analysis of similar products, having as much information as possible on a single product was an important part of being able to decide the pass and fail conditions of the testing processes.

In the interest of keeping focus on the testing process itself, rather than on the relative quality of any particular product that we were testing, we opted to not name the manufacturer, make, or model number of any of the products we tested. Unlike groups such as Consumer Reports that use the standard for comparative analysis to make recommendations within a product category, we used each product in our testing handbook as a stand-in for its larger product class, and a demonstration of how one might test that variety of product. While our results reflect our actual findings, they are only presented as examples of how one generic product's test results may look if tested with our handbook.

What Products did we Ultimately Choose?

The products we selected were a **smart lock**, a **wifi**, and a **connected baby monitor**. We tested them in that order. All three selected products are designed to be used with mobile apps, which is the predominant practice among smart consumer devices. We wanted to make sure that we tested an app component with every product we selected, as so much of the "attack surface" of IoT devices is on these companion apps.

We chose the lock because it was made by a well-known lock maker who now makes app-connected locks through a partnership with app developers, as part of a broader suite of smart-home products such as doorbells and thermostats which are all controlled using the same third-party app. The particular model we tested was also interesting because it does not come with a physical key—one must have a working smart phone to unlock the door from the outside. This raises the stakes of a digital compromise and takeover of the device.

We chose the pressure cooker because of the clear physical safety hazard presented by a rogue pressure cooker. It also represents a company pivoting from the manufacture of solely electrical/digital kitchen appliances to micro computer-driven, internet-controlled electrical appliances. This is an increasingly common pivot for manufacturers as many existing kitchen appliances are turned into smart devices, which made it important that we include at least one such device in our testing handbook. Unlike other categories of IoT, kitchen appliances require some amount of physical interaction. For example, a coffee maker may be able to start the coffee timed to your mobile phone's alarm, but at some point you will have to load coffee and water into the machine. This naturally limits their feature set, and in many cases, the number of other smart products they interact with. The pressure cooker's lack of integration with other products and limited featureset also made it an interesting case study.

We chose the baby monitor because of the clear privacy and security implications that a possible compromise by a malicious attacker represents. These devices gather audio, video, and in some cases health data directly from the bedrooms of infants and toddlers, making this data some of the most sensitive imaginable. The monitor we chose consists of a camera and a mobile app that can relay video and sound data collected by the camera, and also contains its own handheld receiver, which presented a separate, non app-based interface to test.

How did we Design the Technical Testing Procedures?

As we prepared the testing handbook, we realized that some tests included in the standard are much more technically complicated than others. Many of the technical tests require an in-depth understanding of mobile application development and digital security principles, as well as a comprehensive understanding of the features contained within a particular product. These tests also require the set up of custom testing environments to capture certain elements of a product's behavior, as well as specialized software for analyzing code or inspecting captured network traffic. This led us to create processes that in several places suggest more background reading about related topics and software, and may require some higher level of background expertise. While we endeavored to make the handbook readable for the non-expert, there is no getting around the complexity of these tests requiring highly-technical documentation.

In developing the processes for the technical tests, one key area of focus was the ability for a wide variety of stakeholders to conduct testing, even with limited budgets. This requires limiting the scope of certain procedures; when in some cases it was technically possible to dig deeper, we maintained a sensitivity to cost. As researchers at a non-profit, we realized that others interested in similar product evaluation may have limited resources to use on testing. We found that much of the testing required by the Digital Standard can be done without building out a lab full of specialized equipment. The vast majority of the testing processes we describe can be conducted using an average laptop running free and open source tools.

Our testing handbook does require some dedicated hardware for testing—like a wifi network on which you are allowed to inspect all traffic—in addition, of course, to the products being tested. In most cases, a mobile device is required for testing interaction with the product, though depending on how an app is built, it may be possible to run some of the tests using virtual machine emulation instead of a physical device. Of the three products we tested in developing this handbook, only one had good support for emulation.

In some cases, we realized that more information could be obtained if we bought more expensive and specialized testing equipment. For example, while it may be possible to detect activity on a running embedded chip using special clamps designed to read electrical impulses going to the chip’s pins, it seemed out of scope for an achievable and reproducible testing process. We noted these kinds of limitations in our results.

There is a similar limitation to all of the technical tests; if other product testers apply more resources to these tests, they would have to expand their procedures beyond those covered in our handbook. This includes spending more time on things like background research, traffic analysis, and deeper app decompilation efforts, as well as spending more money on equipment. We intend this handbook to be a floor rather than a ceiling, and hope that future testers will expand upon our processes by building up their testing gear or dedicating more time to a particular procedure.

Writing a standardized technical test presented another unique challenge. The relative speed at which change occurs in tech creates a risk of testing for outdated best practices. Old best practices are constantly being swapped out for new ones, and some best practices may still be contentious. For example, recommendations about password strength have steadily moved toward increased password complexity. But how far this should go (and at what point human memory is stretched) is still **an active debate** among security professionals. Those are instances where a tester has to make a qualitative call on the current best practices. We noted where we made such calls in our results.

The Digital Standard's Product Stability test is an example of both the need for qualitative analysis and the danger of documentation growing stale. This test centers on software "fuzzing," the common testing practice of repeatedly providing known bad inputs programmatically into every place within a piece of software that accepts them. These tests are run in loops for days or more, to see under what conditions the code might crash. Fuzzing itself is a far from standardized process, and typically requires deep research of the specific code base being tested. Given the detailed per-product level of knowledge required, and the lack of accepted benchmarks for fuzzing, it is the only section of our testing handbook that is qualitative rather than pass/fail. Our handbook describes one approach to fuzzing. However, **new research** suggests that this process can be automated to a greater degree, and that future fuzzing efforts may well require less specific knowledge of an application's functionality. This will also make it easier to compare results across devices and platforms, as well as make the process less time intensive.

We did not limit our focus on achievability to the cost of equipment. There were tests we knew were technically possible but prohibitively difficult in some other way. In this sense, achievability meant purposefully leaving some possible examinations out of the handbook. To use a previous example, it may be possible to extract microcontroller code from running chips with special equipment; however, in addition to that equipment costing money, it may only be useful for a limited subset of products, and its purchase doesn't guarantee that the code can be meaningfully accessed anyway. For third-party testers, it is generally much harder to obtain access to those parts of a product's running code than it is to get to other parts, like an app. With no guarantee of obtaining code that can be analyzed, it did not seem worthwhile to spend resources on procedures with an unsure outcome.

We also decided to not create processes to test the interaction between devices and "home assistant" technologies. Smart home assistant products present a whole new class of attack surface. They are also mostly devices that run on closed platforms using software controlled by their manufacturer, and not the manufacturer of the IoT products being tested. This makes it harder to fully test or understand product/assistant interaction. While there is some standardization of the communication between IoT devices and home assistants, generally testers would be unable to fully control the test environment of such devices. There is also a relative difficulty in setting up open source implementations of home assistant technology, and in knowing if those implementations are feature complete when compared to more common home assistant platforms. As the landscape of IoT and home assistants changes, it may become more possible to test the interaction between consumer devices and assistants. However, for the time being we felt that in addition to technical complexity, testing with home assistants could introduce uncertainty into the results.

Finally, we chose to focus on testing apps built for Android devices even though all three of the products we tested have both Android and iOS versions of their apps. We made this choice for a number of reasons, but at the top of the list is the size of the Android-user base. While Apple products enjoy wide popularity, far more people are using apps within the Android ecosystem. On average it is also cheaper to purchase testing equipment for Android devices, and as noted above, we endeavored to keep down costs for our testing. The other top reasons are closely related to Android's open source nature: The tool sets are all freely available, and there is a much larger community of developers and researchers working with the platform. We did not make this choice because we don't think that apps for Apple iOS aren't worth testing. Rather, we wanted to focus deeply on the questions asked by the indicators, and we were able to build on the team's existing Android knowledge and the wide array of available documentation on Android systems.

Given all of these factors, the broad range of IoT devices and features, and an understanding that other possible testers may have time constraints, we made decisions in our approach to technical testing procedures that allowed us to prioritize achievability. Often this resulted in our describing the simplest test that would produce a clear result.

How did we Design the Policy Testing Procedures?

The standard also includes a number of tests that evaluate the policies of the product being tested, and those of the company that made or maintains the product. These are designed and function somewhat differently than the more technical tests. In evaluating the indicators for the policy tests, testers mostly review information provided by the company or manufacturer on their website or on documents included when a user purchases the physical product. Some indicators seek yes or no answers to, for example, questions about a product's terms of service or privacy policy.

In some ways it is easier to use a stringent rubric for tests that involve reviewing written documents or using the app itself, because there are fewer functions to evaluate. However, similar to the technical tests, the requirements provided in the standard are often ambiguous. Measuring whether a document is easy to understand or whether data sharing practices are reasonable is completely subject to interpretation by testers. Further, the nature of the product may mean that privacy risks for some products are more significant than for others. It is less crucial that a legal document is “easy to understand” when the product in question poses fewer threats to privacy, for example, if the product does not collect any sensitive personal information. However it is extremely important that a product which collects biometric, location, and other sensitive data has a privacy policy that is understandable by users, because mishandling this data can have devastating effects to a user’s privacy and security.

In designing the procedures for these tests we often focused on whether the company provided clear examples or definitions in their policies, and whether the policies addressed all of the functions of a specific product. Other tests focus on the specific practices of the company—for example how they manage specific user data, or what internal company practices on privacy are. A company may have a robust data deletion policy, but unless they specify that they do in documents that are available to customers, we were forced to fail them on relevant indicators. This is both a limitation of the standard—in that results may not always reflect actual practices—and a strength, in that it is arguably more important that a user has an accurate understanding of a company’s practices so that they can make educated choices. Even when companies do appropriately delete or minimize data according to best practices, it is still important that they disclose that information to users.

The policy tests are likely more accessible to non-expert testers than the technical tests, which was why it was crucial for us to make sure that the processes were clear and consistent. Anyone can read company documents or use the product and see what its features are, whereas not everyone can evaluate code to look for vulnerabilities. Given this, we did our best to create procedures that could be replicated by anyone and did not require significant interpretation and expertise.

What would we Change in the Standard?

Developing our testing handbook and testing the three selected products allowed us to make observations about the standard, and consider ways that it could be refined and improved. As it is an open source project, contributors are encouraged to suggest changes and updates to the text. These are seven recommendations based on our experience.

1. The Digital Standard should provide more context as to what the tests mean and why each best practice is important for protecting privacy and security.

As it stands, the Digital Standard tests allow testers to gather results and answer questions, but do not provide testers the necessary tools to interpret those results in the broader privacy and security context. Each best practice that the standard tries to incentivize is important for a different reason, not all of which will be apparent or familiar to all potential testers or users of the standard. For example, establishing whether a company publishes a regular transparency report regarding a product is less helpful if testers or readers do not understand the value of or rationale for transparency reporting. Further, evaluating the data retention practices of a product will be more informative and meaningful if testers understand how data retention or minimization can affect privacy and security.

Although some of these best practices are relatively straightforward, many are not. If a goal of the standard is to enable third-party organizations or individuals to test their own products, then we need to make sure that they have the necessary understanding of the relevant 35 areas of digital security and privacy. Providing context and links to best practices would make the Standard more effective and informative, especially because one of the goals is to help consumers make informed decisions about the products they use or purchase.

Some of the tests may be more relevant than others, depending on the product being tested. For example, it is probably less crucial that a product that obviously collects very little personal data provides as comprehensive a set of third-party request policies than one that collects detailed biometric and location data. Passing or failing that test has different consequences for different products. Understanding the context of the tests, and the best practices they evaluate, makes the results of those tests more meaningful and valuable.

2. The Digital Standard should include guidelines for how tests are meant to be scored.

The Digital Standard, as written, does not provide any information about how to score tests, instead offering criteria and indicators. Our initial approach was a simple Pass/Fail system, but we quickly realized the need for more nuance, and included a Partial Pass grade as well as a Not Applicable grade to broaden our range of options. For example, a product that does not collect any personal information in the course of its operation would otherwise get a failing grade when evaluated on whether it gives users the ability to control what information is collected, so we might mark it as Not Applicable. But even our solution is unsatisfying insofar as it doesn't positively reward the product for not collecting user data in the first place, acquiring no data for users to control.

In order for our processes to be replicable by other testers, we needed to develop clear conditions for Pass, Partial Pass, Fail, and Not Applicable results, so that anyone using the handbook would be working from the same set of rules. A more nuanced alternative could have been a numerical score, but that proves even more subjective without a ridgid rubric for each score. What one tester may consider average is different from another, and this proves even more subjective due to the varying expertise of testers who may have different standards for success or failure. Although our resulting scoring system was the clearest process we could come up with, the standard itself should provide official guidance on how to score tests and indicators to insure replicability and trust in the results.

3. Components of the Digital Standard should be weighted according to the priority of the tests and indicators when actually measuring the impact of these practices on privacy and security.

Not all of the 35 tests included in the Digital Standard are equally important in evaluating how a product would perform in protecting a user's privacy and security. At the present time they are not weighted or prioritized, even though some are clearly more important and crucial than others. For example, whether a user has the right to repair a product themselves is not weighted any differently than whether that product uses strong authentication practices. This lack of weighting or prioritization can inaccurately reflect the quality of a device should, for example, it pass all of the crucial tests but fail some of the more obscure ones.

Similarly, tests that include multiple parts do not prioritize individual indicators for evaluating whether a product has passed or failed. We attempted to reflect this through partial pass results, but this is a limitation of the Digital Standard more broadly. For example, an indicator measuring whether a company discloses the kind of encryption they implement is not as important as an indicator measuring whether encryption is actually used or not, but both are considered necessary to pass the encryption test. This kind of revision could also address the questions of dependencies between indicators. For example, if a company does not produce a transparency report, then the product cannot be tested on any subsequent indicators that require reviewing a transparency report.

A hierarchy of importance in both tests and indicators could also be useful in signposting for testers and consumers of the Digital Standard which of its many indicators to focus on first; which ones are vital, and which are helpful, but not critical.

4. Tests and indicators need to be constantly updated due to the changing nature of digital security best practices. They also need to provide internal flexibility to adhere to best practices for specific products.

While conducting our own testing, we noted a number of indicators for which the conditions for a product's pass grade will have to change over time due to advances in technologies or best practices. Technology is changing so quickly that what may be a best practice in 2020 could easily change in the future. For example, one indicator calls for services to require passwords that are at least eight characters long, which was set at the last time the standard was updated. Best practices for password complexity are constantly changing, for instance by increasing the minimum required number of characters or requiring the addition of various capitalizations, special characters, or numbers. The Digital Standard should try to build in some kind of future planning so that it stays flexible, either by explicitly sunsetting some indicators and requiring the Digital Standard contributors to decide whether to update the requirement, or by pointing to some external standard that does shift with evolving reality, such as those published by National Institute of Standards and Technology (NIST).

Relatedly, there were a few instances where we had trouble rating a product on an indicator when the product was actually too good and implemented best practices for privacy and security that were not yet included in the standard. For example, the baby monitor we tested failed a strict interpretation of the two-factor authentication test, however the handset and device implemented other safety features that served as factors that limited the ability for harm. While not second factors in the strict sense, the monitor achieved a similar security benefit by requiring both the handset and the device to operate on the same local secured Wi-Fi network in order to communicate. This gap is partially because the standard is not regularly updated to include current best practices, and partially because best practices vary so much from product to product that it is hard to capture them in an evaluation system that is supposed to apply to a variety of product verticals. This problem could be addressed through an improved scoring system that, for example, allows adjustments to final ratings using some kind of extra credit measure or compensating overall score for specific practices.

Alternatively, extra indicators could be added that allow for the recognition of supplemental best practices beyond the scope of the existing Standard criteria.

5. Tests that require analysis of a product's legal and policy documents (specifically privacy and terms of service/use policies) need to be more specific about what best practices are required under the Digital Standard, perhaps differentiating by product category or capability.

We found that our analysis of legal documents relating to products that we tested was hampered by a lack of clarity in best practices in these areas (particularly as they pertain to smart devices). Products collect different types of information, or offer different features, that make it hard to create a scorable template for an ideal privacy policy. For example, a product that does not collect sensitive information is still required to include processes for law enforcement requests of different types and from different jurisdictions. A Not Applicable rating requires the tester to subjectively decide whether a specific indicator is important and relevant for a specific product. Alternatively, requiring every product to include all indicators in their privacy and terms of service policies could either penalize products for failing to provide unnecessary information, or make the policies much more confusing than they need to be by forcing the inclusion of irrelevant information.

In its technical tests the standard already differentiates the processes for testing by specific features of the product being tested. For example, the Known Exploit Resistance test provides different procedures for products that use browsers, apps, and connected devices. A product may use two or more of these things, like the baby monitor we tested which consists of a mobile app and a handset. Similar differentiation could be applied in the policy sections of the standard as well. These could be differentiated by the type of data the product collects, for example one that collects personally identifiable information versus one that does not, or some other method of identifying whether a product poses a higher privacy risk.

Deciding what the ideal policies for a specific product should look like requires expertise in that product vertical, relevant regulations, and for the tester to make subjective decisions about what best practices should be expected from a given product. This makes testing inaccessible to non-experts and does not ensure that different products will be tested using a similar rubric. There are debates among experts about the virtues of simpler, more limited privacy and terms of use policies, which may be more accessible to a reader. Shorter privacy policies are also more likely to be reviewed by customers, compared with longer and more comprehensive policies that may address every possible best practice, but might be discouraging based on length.

The standard needs to make more explicit its preferred best practices, potentially providing product or capability-specific guidance. It also needs to clearly define terms like “understandable” to allow for replicable and consistent testing.

6. The Digital Standard needs to address the fact that companies’ legal documents are often convoluted or vague due to complicated product lines, partnerships, and supply chains.

We had challenges evaluating the legal commitments that companies expressed in their terms of service and privacy policies because of the sometimes complicated connection between the legal terms and the products being evaluated, and because of situations where multiple companies have different legal obligations all covering the same product. Companies that have many different services and products, for example, may have documents that are not explicit about which products are covered by which policies. Similarly, it can be hard to distinguish between policies that are meant to apply to physical products and those which are only intended to cover a web site. Companies that have many products can also present difficulty, as it can be hard to be sure which policies apply to the product under evaluation. This mess of legal documents isn’t a flaw of the Digital Standard, of course, but there are some indicators that would benefit from more clarity about what it expects from companies and how detailed the language it is looking for should be.

In other circumstances we observed issues that privacy concerns arose because a product being evaluated included features from another company. For example, if a product interfaces with a home assistant platform, that interface may have impacts on privacy that are not addressed in the manufacturer’s policies. It may also mean that policies apply to specific aspects or features of a product, but not others. The IoT is a connected web of products which, by design, interface with other products or services. The complexity of these relationships and connections means that best practices for evaluating a product’s legal documents need to provide more guidance to ensure that testers are consulting the correct material and accurately placing responsibility for the privacy and security of users’ data.

7. The Digital Standard needs to address the fact that, as written, some indicators cannot be accurately tested.

There are some indicators in the Digital Standard that can only be evaluated by observing the device being tested and making educated guesses as to how the product is behaving. It is nearly impossible to arrive at a confident statement as to whether a product passes or fails. The best reviewers could do is to make observational assumptions. For example, one indicator regarding data use reads, “The company explicitly discloses every way in which it uses my data.” Finding the ways that the company states they will use the data is usually as easy as reading a privacy policy. However, an outside observer seeking to determine how they actually use the data, can only make educated guesses by watching the product and looking for behavior that would demonstrate some data use not mentioned in the policy.

Conclusion

The Digital Standard is extremely valuable as a consumer-focused tool for evaluating IoT privacy and security. Its attempt to encourage manufacturers to produce better products, while educating users about the devices they use, is unique among security standards, as is its open-source nature, and the opportunities it offers for external actors to contribute to developing and maturing the standard itself. Through testing various products, and the design of our own testing handbook, we developed this set of recommendations and appreciate this opportunity to engage in the standards-setting process. The aim of this project is to create tools that will help other organizations, researchers, and consumers conduct their own testing, and we make our seven recommendations with that goal in mind.