Developing a Criteria-Based Evaluation Tool for User Experience Design that Balances Standardization and Creativity

Erin Hopkins
Systems Engineering
University of Virginia
Charlottesville, VA
edh5ps@virginia.edu

Jacqueline Mazzeo
Systems Engineering
University of Virginia
Charlottesville, VA
jpm3ay@virginia.edu

Vinh Nguyen
Systems Engineering
University of Virginia
Charlottesville, VA
vqn2ed@virginia.edu

Emma Peck
Systems Engineering
University of Virginia
Charlottesville, VA
esp2kq@virginia.edu

Kelcie Satterthwaite
Systems Engineering
University of Virginia
Charlottesville, VA
kls9aw@virginia.edu

Carlos Lidón
King Digital Entertainment
Barcelona, Spain
carlos.lidon@king.com

Gregory J. Gerling
Systems Engineering
University of Virginia
Charlottesville, VA
gg7h@virginia.edu

***Abstract* – Design systems are increasing in popularity, created to ensure a consistent aesthetic of graphics and interactions in websites and apps, and guide product development. They tend to include a set of standards, principles, and documentation. Due to their often-rigid requirements on structure and uniformity, traditional design systems can discourage creativity and customization. To forge a balance, the work develops a criteria-based evaluation tool, or 'scorecard' for assessing design components that incorporate principles of consistent, standardized practice, yet prioritize creative freedom. The evaluation scorecard allows inconsistencies to be managed in a collaborative and consensus-based manner. Users select parameters and metrics to evaluate the various elements of a design component. The tool calculates a score based on the number of parameters passed or failed. Scores below team-desired thresholds signal a need for further modification or redesign. Usability feedback using talk-aloud and surveys in a focus group format assess the ease of use and efficiency of the tool and identify gaps in functionality.**

Keywords—user experience design, user interface, design system, standardization, usability

I. Introduction

Design systems are being developed and used increasingly in the field of user interface/user experience (UI/UX) design to facilitate the development of products, such as websites and apps. Indeed, many large companies have developed their own design systems, including Google, Microsoft, Apple, and IBM. Their design systems are not "one size fits all," but offer insights into the orientation of an organization. Such design systems characterize the purpose and shared values of a product with groups of designers who often work semi-independently [1]. Established design systems holistically organize a consistent set of specifications for the product. They help facilitate a standardized aesthetic [2]. The absence of a design system can lead to inconsistencies and stylistic flaws relevant to icons, buttons, text style, language, navigational sequences, etc.; and consequently, a confusing and/or inconsistent end-user experience [3].

On the other hand, traditional design systems can restrict the artistic freedom of UI/UX designers by confining their design choices. Due to such restrictions, in certain instances and organizations, highly innovative designers opt to preserve their creative autonomy and forgo the use of a design system. Conversely, when creativity becomes overly prioritized, designers can lose focus on how their artistic nuances fit into a coherently designed product. Thus, there is a need to balance standardization and customizability.

Herein, we describe a criteria-based evaluation applied to design components to afford creative autonomy while maintaining some level of consistency. The criteria-based tool takes the form of an evaluation scorecard, which allows designers to assess the elements within a design component and their adherence to shared design standards agreed upon by the team. The scorecard tool is applicable at various stages of design, from initial planning through launch and post-production. Designers can evaluate many different design elements, including icons, or pop-up sequences. They might evaluate a design that is newly created or update one pre-existing. In so doing, the designer formally characterizes multiple parameters, such as size, color, orientation, placement, and position of applicable elements. That design component's adherence to those specifications per element is evaluated on a pass/fail basis, which leads either to changes to its design or to changes in its scorecard's parameters and metrics.

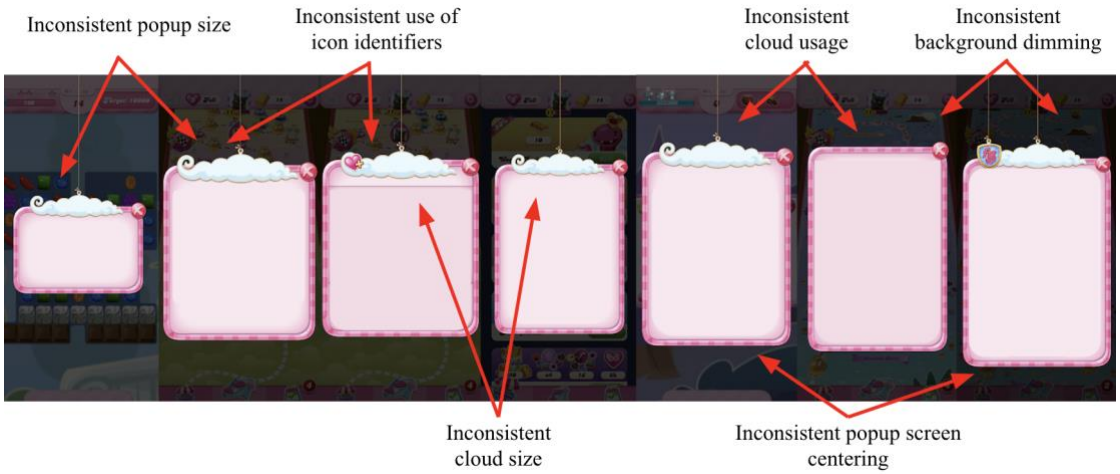

Fig. 1. Inconsistent pop-up design apparent in a mobile game interface.

Fig. 2. Button placement inconsistencies catalogued from a mobile game interface.

II. Methods: Delineating Design Inconsistencies

To better understand inconsistencies that arise absent a design system, a case study was performed of the game Candy Crush Saga (King Digital Entertainment). Delineating design inconsistencies is a first step in moving toward coherence and consistency. Inconsistencies were grouped into the four general categories of representation, presentation, navigation, and interaction [4]. In brief, representation refers to which visual elements are chosen to represent important informational relationships. Presentation refers to ways visual elements are organized and laid out to relate to each screen. Interaction and navigation regard ways in which users interact with dynamic fields and transition between screens.

Inconsistencies with regard to representation include button design and inconsistent pop-up windows, Figs. 1-2, as well as typography and color palette. Moreover, issues with presentation largely center around the placement, arrangement, and position of buttons. Exit buttons, for example, are laid out in different locations throughout various stages of the game, Fig. 2. Inconsistencies in interaction include differing fail-level sequences and false affordances. Inconsistencies in navigation encompass issues with map navigation and functional proximity.

III. Methods: Design and Implementation

The criteria-based evaluation tool described herein seeks to provide designers with a resource that balances organizational coherence and consistency with individual flexibility and creativity. The tool – organized as a scorecard – is used to evaluate design components based on customized specifications. Terminology to be described includes design components, elements, parameters, metrics, and specifications. A design component is the entire unit or image being assessed. Elements are the individual design aspects that make up a component. For example, a pop-up window design component may include elements of icons, action buttons, text, etc. A list of parameters is used to specify which characteristics of the specified element of a design component are being evaluated. Example parameters include shape, size, and color. Each parameter has one or multiple metrics that provide a means to measure the parameter, such as degree of rotation or font size. Using the parameter-metric pairs, a user specifies the characteristics a design element should have in the specification column. Following this, a user can set a weight for each specification and then evaluate the element by checking if it passes or fails these specifications. The scorecard generates a score per element based on the weights of each specification

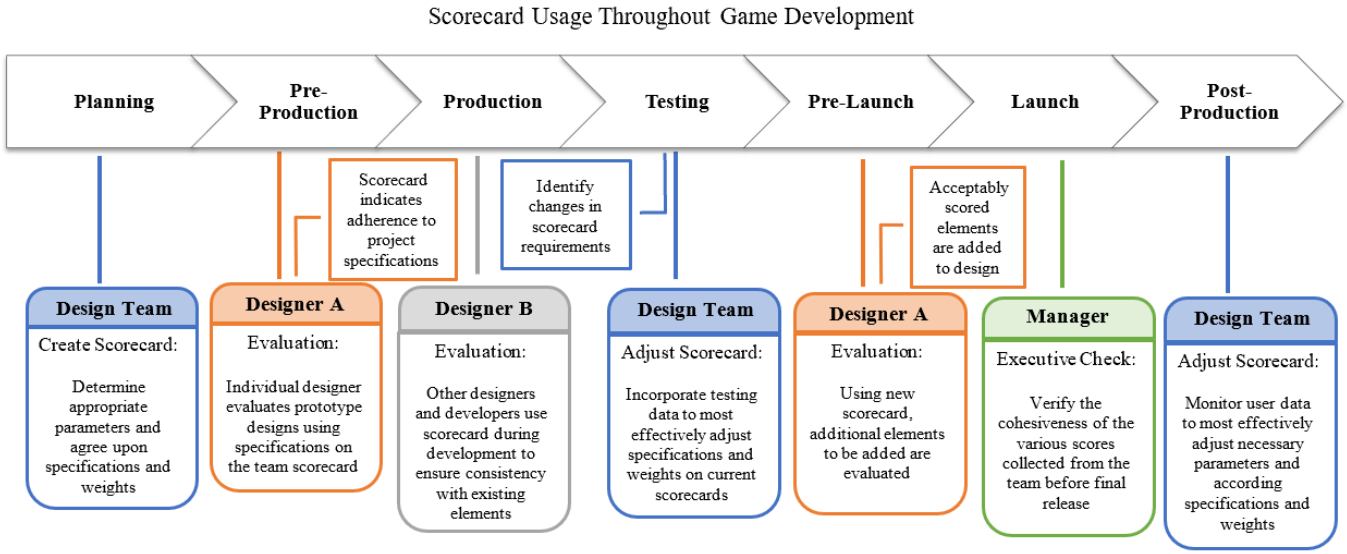

Fig. 3. Process flow diagram of the scorecard tool's integration into the game development process.

and the element's performance against the specified criteria. The parameters and metrics may be rigid, but the designer inherits freedom from the specification they choose and the weight they provide for each parameter-metric pair. Note the process leading up to the final design of the scorecard included multiple iterations and usability evaluations.

Concept and Design

The evaluation tool can be used across stages of design, beginning with planning and continuing throughout post-production [5]. Figure 3 demonstrates the envisioned process flow in which multiple users interact with the tool collaboratively and individually.

Figure 4 demonstrates the use of the scorecard itself, in evaluating the example design element of a complex icon. As the actual parameters and metrics are built into the scorecard, the user inputs neither their values nor their definitions. A singular score is calculated on the top left. The calculation of the score, as well as key features of the scorecard tool are discussed in the following paragraphs. Note that the use of parameters and metrics represent the concept of standardization. The specification and weight column illustrate the concept of customizability.

Parameter. The first decision made by the user is whether the parameter is relevant or irrelevant to the selected element. By unchecking the checkbox to the left of the parameter, Figs. 4-5, an entire row can be hidden and excluded. Parameters include orientation, shape, size, color, animation, dimension, placement/position, text/language, feedback, sequence and navigational gestures.

Metric. Each parameter has one or more defining metric that provide a technical description. For example, the metric for the placement/position parameter is the quadrant of the screen. Another example is the parameter text/language, with four separate metrics: number of words, font style, font size, font color. These are predetermined ways to define the parameter. All users implementing evaluation or feedback with the scorecard tool view the same parameters and corresponding metrics.

Specification. A design team decides a range of values that specify the parameter-metric pair. Examples being the text size or color hex value they will allow to pass or fail per interface element. By allowing designers to specify a range, the elements being evaluated need not be identical but can exhibit creative differences. Additionally, specification cells contain drop-down menus containing suggested options. To encourage flexibility, users can clear the drop-down menu and manually input values into specification cells.

Weight. A 'rigid' weight equates to a value of 1, meaning that this parameter-metric pair is counted equally into the calculation formula. A 'flexible' weight equates to a value of 0.5, meaning a pass or fail will have 50% of the impact on the overall score compared to a weight marked as 'rigid.' This functionality was included because some specifications may be somewhat important to the design, but not required of the design. Weight values allow designers to evaluate designs based on metrics that are not essential enough to fail the entire evaluation (a score <60% or value set by users).

Evaluation. When the user is evaluating a selected component along preset criteria, she or he will decide whether its composing elements pass or fail each specification. To provide visual cues, the tool includes conditional formatting that turns red for design specifications that fail, and green for specifications that pass.

Element:	Piggy Bank
Score:	66.7%
Parameters Evaluated:	9

	Parameter	Metric	Specification	Weight	Evaluation
☑	Orientation	Tilt	Left tilt	Rigid (1)	Fails
☑	Orientation	Degree of Rotation	45°	Flexible (.5)	Fails
☑	Shape	Shape type	Circular	Rigid (1)	Passes
☑	Color	Color Contrast Ratio	3:1	Flexible (.5)	Passes
☑	Color	Hex value	Yellow	Rigid (1)	Passes
☑	Dimension	Number of dimensions	1D	Rigid (1)	Fails
☑	Text/Language	Number of words	1-5	Flexible (.5)	Passes
☑	Text/Language	Font Style	Candy Crush Font	Rigid (1)	Passes
☑	Text/Language	Font color hex	Brown	Rigid (1)	Passes

Fig. 4. Design evaluation of a static icon using the scorecard tool. Users performing an evaluation of the icon on the right verified that nine parameter-metric pairs are applicable to this design component. Users also recorded the specifications for these parameter-metric pairs in the Specifications column. As denoted in the weight column, six metrics were identified to be rigid and the remaining three are flexible. When evaluating based on their design criteria, three parameter-metric pairs failed. Accordingly, the design element being evaluated received an overall score of 66.7%.

Element:	Pop-up
Score:	90.9%
Parameters Evaluated:	7

	Parameter	Metric	Specification	Weight	Evaluation
☑	Shape	Shape type	Rectangular	Rigid (1)	Passes
☑	Size	Height, width	Width: 80%	Flexible (.5)	Fails
☑	Color	Hex value	#FCE5EF	Flexible (.5)	Passes
☑	Dimension	Number of dimensions	2D	Rigid (1)	Passes
☑	Placement/Position	Quandrants of screen	Center	Rigid (1)	Passes
☑	Exit Navigational Gesture	Motion to exit	Tap X button	Flexible (.5)	Passes
☑	Continue Navigational Gesture	Motion to get to next screen	Tap forward action t	Rigid (1)	Passes

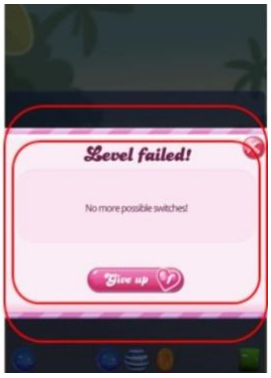

Fig. 5. Design evaluation of a pop-up window using the scorecard tool.

Score. The scorecard tool calculates an overall score out of 100. The score indicates the adherence of a design to criteria specified in the scorecard. The score takes into account how many parameters are evaluated, the weight of each parameter-metric pair, and how many specifications pass and fail. The score is a weighted average calculated using the number of parameters selected; a failure is a 0% and a pass is a 100%, each weighted 1 or 0.5 based on the assigned weight value.

How to. A metrics guide was created for the user to better decipher and understand each parameter and metric pair. This resource is integrated into the tool and remains available to the user throughout the scorecard's lifecycle.

IV. Methods: Usability Evaluation

Two usability evaluation sessions were conducted with the scorecard tool to evaluate its utility, understandability, and ability to balance standardization and customizability. Four usability experts participated in the first session and thirteen domain-specific designers participated in the second session. The selected use cases addressed the tool's features of parameter-metric pair specifications, weight selection, and design evaluation.

To familiarize participants with the functionality of the scorecard tool, participants were guided through introductory activities that 1) overviewed a blank scorecard, highlighting its main features, 2) walked them through a completed scorecard, and 3) asked them to evaluate passing or failing specifications given an element and a scorecard with pre-selected parameters, metrics, specifications, and weights. Participants were then split into two groups to complete the use cases. The time taken to complete each activity was recorded.

A. Use Case 1a: Metric Specification

Using an early instantiation of the scorecard, usability experts were provided a pop-up design component, Fig. 6A. In this earlier instantiation, users could not decompose a design component into smaller, sub-elements. Users were tasked with approving the scorecard tool's pre-selected parameters, as well

Fig. 6. Mobile screenshots used in usability evaluations in A) Use Case 1a, B) Use Case 1b, C) Use Case 2, Session 1, and D) Use Case 2, Session 2.

as identifying specifications given characteristics of the pop-up design component.

B. Use Case 1b: Metric and Weight Specification

Domain-specific designers were presented with a pop-up design component broken down into its elements of a pop-up body, exit-button, action button, and score value, Fig. 6B. Each element had a designated scorecard with pre-selected parameters. Users were tasked with identifying appropriate metric specifications based on the characteristics of the pop-up body element.

C. Use Case 2: Evaluation using Scorecard

In the final usability evaluation activity, participants in both sessions used the scorecard with the specifications determined in use case 1 to evaluate the designs of an alternative pop-up. Usability experts who previously specified the scorecard according to Fig. 6A in use case 1a evaluated Fig. 6C, while the domain-specific experts that specified metrics based on Fig. 6B evaluated Fig. 6D. Subsequently, participants gave general feedback on the scorecard tool both verbally and via a written form consisting of open ended and Likert scale questions.

V. Results and Discussion

This work created a novel criteria-based evaluation tool – in the form of a scorecard – to be used for assessing UI/UX design components. The use of the scorecard, in contrast to traditional design systems, seeks a balance between artistic creativity and standardization. In the domain of gaming considered herein, artistic creativity is significant and non-standardized, which was evident in differences in how the scorecard tool was completed by separate groups in usability evaluations. On the other hand, the tool's ability to create a level of standardization was apparent in the collaboration and agreement that took place. The separate groups of participants were able to collaboratively generate and agree upon design guidelines in their assessments of pop-up screens. Finally, participants' feedback notes they found it to be easy to use and practically useful.

The usability evaluation sessions provided valuable feedback on the scorecard evaluation tool, Figs. 7-8, and Table 1. The responses were generally positive. Some suggestions were that element sizes could be assessed as their percentage of screen space, background padding could be added as a parameter, the distance between words and other graphical elements could be included, and color contrast could be augmented with regard to accessibility and color blindness. There was also an articulation of the value of the tool and its helpfulness for younger designers or product managers, as it articulates specific design criteria. Participants believed the scorecard tool would be used when introducing a new component to an existing screen, to ensure it matches the present style.

The usability evaluation also considered performance differences between separate groups evaluating the same interface. We found that groups in both sessions differed trivially in the parameters selected, specifications decided, and weights chosen. This finding indicates that the scorecard tool represents a shared UI/UX language, with potential to create and enforce guidelines on component representation and interaction. In the sessions, participants worked to achieve a standard and ideal outcome upon which they all agreed. They also articulated the importance of having standardized components, which may benefit future design thinking, in addition to addressing the current issues and inconsistencies.

We evaluated the length of time spent on each evaluation activity. To complete Use Case 1a, Group 1 took 5 minutes and 30 seconds and Group 2 took 12 minutes. For Use Case 2, Session 1, Group 1 took 3 minutes and 45 seconds and reached an evaluation score of 83.3%, while Group 2 took 7 minutes and 21 seconds and reached an evaluation score of 66.7%. The range of about 4 to 12 minutes to complete the evaluation seems reasonable to ask of a design team introducing a new component. Further, the differences in time spent exemplifies

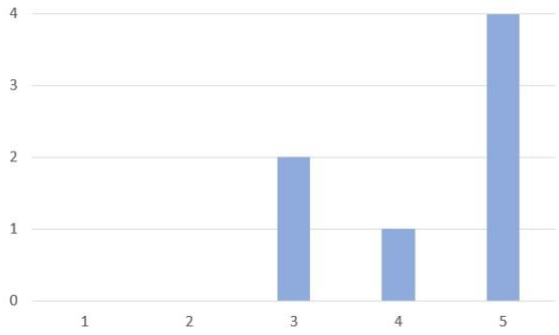

Fig. 7. Survey results showing participants' ratings of how helpful was the scorecard tool (1, not very helpful; 5, very helpful).

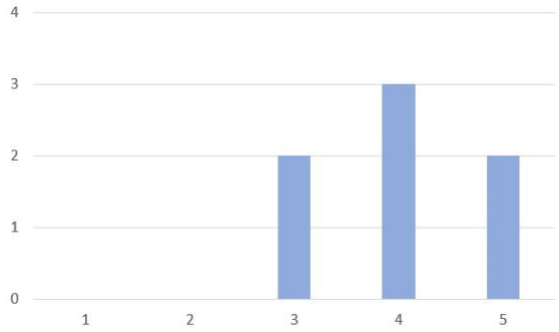

Fig. 8. Survey results showing participants' rating of how easy to use was the scorecard tool (1, very difficult; 5, very easy).

Table 1. Participant responses to feedback survey.

Survey question	Summary of participant responses
What are some things you like about the tool? Is this something you would use?	- Once it is prepared, it would be a very efficient way to check for consistency - It is super configurable and super easy to use, and has a lot of possibilities - Great tool to quickly and efficiently spot issues or improvement in design - Liked the amount of options of parameters and metrics for evaluations - Sees value in utilizing this tool in workflow since it is easy to use
What are some things you dislike about it?	- Exceptions have to be considered - The percentage score could be changed to "pass vs rework", can be decided by users - More breakdown options for different layouts focusing on logic
Do you have any recommendations on how to improve it?	- Continue user testing and iterating the tool based off feedback and findings - Parameters could be labeled in a manner more familiar with the industry - Make more room for comments - User testing with end users could find more areas of improvement

how the tool generates varying levels of discussion, which is indicative of user influence and control.

The difference in score evaluations, as aforementioned, highlighted a need for more standardization within the scorecard tool in an early instantiation. Indeed, the two groups' scorecards differed in parameter selection. Group 1 selected 7 parameters and Group 2 selected 8 parameters. The additional parameter was orientation, which was given a 0° specification. The impact of participants specifying a 0° orientation versus not including orientation as a parameter is extremely minimal. Another specification that differed was size. One group gave a range of 50-60% while the other specified 80%. A further difference was with the weights participants selected per specification. Both groups indicated that parameters of shape and continue navigation were rigid and the parameter of size was flexible, but had different weights for the parameters of color, dimension, placement/position, and exit navigational gesture. These minor differences exhibit the tool's ability to cater to user preferences. While weights can be subjective to an individual, the group of designers can agree on a standard that will reduce inconsistencies in the long-term. Moreover, as most of the differences were found in the weight category (rigid, flexible), this highlights that the tool caters to customizability, while the minor distinctions in parameters shows that the tool is usable and uniform.

Finally, due to the on-going SARS-CoV-2 pandemic, the evaluation usability sessions were completed in an on-line setting. This may have hindered discussion due to participants feeling more hesitant to unmute and speak up. Additionally, due this setting, the sessions may have not captured other interpersonal aspects. An important step in the future is to evaluate the tool in a physical meeting.

ACKNOWLEDGMENTS

We would like to acknowledge the advice, support, and feedback of personnel at King Digital Entertainment. We also acknowledge Chris Grant for his guidance and support.

REFERENCES

[1] Hacq, A. (2020). *Everything you need to know about Design Systems*, UX Collective. Retrieved March 30 2021, from https://uxdesign.cc/everything-you-need-to-know-about-design-systems-54b109851969

[2] *Material Design*. Retrieved April 1, 2021, from Google https://material.io/design.

[3] *Carbon Design System*. (2020). Retrieved April 1, 2021, from IBM Open Source https://www.carbondesignsystem.com/.

[4] Spence, R. (2014). *Information Visualization: An Introduction 2014 Edition* (3rd ed.). Springer Publishing.

[5] Pickell, D. (2019). *The 7 Stages of Game Development*, Retrieved April 02, 2021 from https://learn.g2.com/stages-of-game-development