11. Contract design beyond the hype: measuring the value

Marie Potel-Saville and Mathilde François Da Rocha

11.1 INTRODUCTION

This chapter focuses on studying the value created by contract design: how to measure it to constantly improve it, leveraging scientifically grounded methodologies to enable the sharing of results between practitioners.

Basically, contract design practitioners agree on the fact that the point is no longer *only* legal expertise, but also human factors and the acceptance of contracts. This touches upon the purpose of contracts: status or function contracts, as explained by Weber or the five dimensions of contract according to Bellay[1] (the insurance-policy contract, the bureaucratic, the standard, the community and the moral contract).[2]

Contract design advocates an evolution of the practice of law from legally-centered to user- or human-centered. Among the many reasons for this, efficiency ranks high on the list.

In a research report on the purpose of contracts,[3] the International Association for Contract and Commercial Management (IACCM) identified that on average, contracts underperform against expectations by 27%. In particular, there are 2 main areas where contracts short fall of expectations:

1. providing the information people need to do their job (74% of respondents consider that contracts do not perform this function well), and
2. focusing negotiations on topics needed to achieve mutual success (70% of respondents indicated that contracts do not perform this function well).

Besides, mere compliance of contracts with some areas of the law trigger a need to create more understandable and accessible contracts.

[1] Jean-Guy Belley, 'Max Weber et la Théorie du Droit des Contrats' ('Max Weber and the Theory of Contract Law') (1988) 9 Droit et Société (Law and Society) 281.

[2] For an analysis of the categories and how a designer would look at them, see Marie Potel-Saville, 'Shaping the Law to Restore its Function?' (Professional thesis at ENSCI, Innovation by Design Master's Degree, January 2020).

[3] IACCM, 'The Purpose of a Contract – Research Report' (2017) [https://www.worldcc.com/Resources/Content-Hub/View/ArticleId/8673/The-Purpose-of-a-Contract-An-IACCM-research-report] accessed 23 April 2021.

11.1.1 What Do We Already Know About the Value of Contract Design?

The tip of the iceberg is the 'hype' factor. For example, in February 2021, the *Financial Times* explained how 'design thinking' helps lawyers to do a better job, mentioning an ambitious contract design project by HSBC.[4]

More concretely, the main benefits which have been identified so far include:

- More compliance, ie limitation of legal risks in the numerous areas of law which now require a higher clarity and accessible standard of legal information: privacy (Art 12 of the GDPR), anti-corruption, insurance (Directive 2016/97 of 20 January 2016 relating to Insurance Product Information Document), telecoms (EU Regulation 2019/1150 of 20 June 2019), consumer protection (article 5 of Directive 93/13/EEC);
- Advantage in litigation, for example the UK High Court rendered a judgement in July 2020[5], in which a boilerplate clause contradicted an appendix which had been designed and visualized. Interestingly, the High Court considered that the real intention of the parties was to be found in the 'worked examples' (ie the designed appendix) rather than the boiler plate, because, the role of the worked examples was to evidence the possible consequences in various situations;
- Generating higher engagement by users: for example, displaying the reading time upfront has been proven by the Behavioral Insights Team[6] as multiplying engagement, ie users opening the document online by 105%;
- Fostering understanding, thus implementation by users;[7] diagrams for example enable users to better understand the steps required to complete a procedure, and the correct sequence of required actions;
- Displaying titles as FAQs has been proven by the Behavioral Insights Team[8] as enhancing understanding by 36%, the use of icons by 34%, showing text in a scrollable text box by 26% and illustrations by 24%;
- Supporting attention and information search, deductive reasoning, comprehension and problem solving and recall;
- Bridging cultural, language and literacy gaps between legal experts and clients, jurors, and decision-makers;
- Creating a better user experience.

---

[4] Reena Sengupta, 'How "Design-Thinking" Can Help Lawyers Do a Better Job' *The Financial Times* (London, 11 February 2021).

[5] [2020] EWHC 1891 (Comm) [https://www.bailii.org/ew/cases/EWHC/Comm/2020/1891.html] accessed 23 April 2021.

[6] The Behavioral Insights Team, 'Improving Customer Understanding of Contractual Terms and Privacy Policies: Evidence-Based Actions for Businesses' (18 July 2019) [https://www.bi.team/publications/improving-consumer-understanding-of-contractual-terms-and-privacy-policies-evidence-based-actions-for-businesses/] accessed 23 April 2021.

[7] Stefania Passera, 'Flowcharts, Swimlanes, and Timelines: Alternatives to Prose in Communicating Legal–Bureaucratic Instructions to Civil Servants' (2018) 32 Journal of Business and Technical Communication 229.

[8] ibid.

11.1.2 Where Do We Go from Here?

To advance the practice of contract design, we need a clear methodology, with precise criteria to validate what works, to compare precisely two prototypes and identify what works best and how to quantify the gain, to involve directly users and measure what is the most usable and easily accepted by them.

This means that user-centered evaluation, ie evaluation based on the users' perspective, is a required activity in contract design. How to apply and adapt internationally recognized user-centered evaluation methods to redesigned contracts? How to share results and best practice so as to contribute to reaching systemic impact?

We hope this chapter will provide concrete tools and method to practitioners, tested in the field with concrete use cases. The chapter is structured as follows: After summarizing what we already know about the value of contract design, Section 2 addresses when the evaluation should take place, distinguishing between diagnostic, formative and summative evaluations. Section 3 discusses what to evaluate. It provides an evaluation framework developed specifically to assess contracts and other legal documents, with a focus on acceptability and acceptance and their underlying models, as well usability. In this section, we provide real examples of projects which have been assessed, along with the results. Section 4 delves into details as to how to evaluate, distinguishing between expert audits and user testing, and provides both methodology and concrete examples in some of the projects we delivered. Finally, the conclusion in Section 5 discusses the limitations of the assessment framework so far and further research required.

11.2 WHEN TO EVALUATE?

The assessment of the quality of legal documents can occur at different steps in the design process. The human-centered design process has been described in ISO 9241-210: 2019[9] and consists of 4 iterative phases:

1. Understanding and specifying the context of use
   eg in our case, describing who will use the contract, for which tasks, in which environment or organization
2. Specifying the requirements
   eg the user shall be able to read and understand the contract quickly (<5 minutes)
3. Producing design solutions
   eg several contract designs
4. Evaluating the design
   eg invite several end-users to read the redesigned contract and collect their feedback as well as measures of efficiency

[9] ISO 9241-210:2019, 'Ergonomics of Human-System Interaction — Part 210: Human-Centred Design for Interactive Systems' (2019).

During these different phases, evaluation can answer several questions and has several roles:

- To obtain data in order to define and prioritize the aspects to be improved;
- To evaluate the design with users to improve it and enhance solutions to their pain points based on their feedback;
- Present a prototype during the development phase to minimize the risk that the new document does not meet the needs of users or the organization concerned;
- Discover or confirm some requirements or constraints which could have remained implicit or difficult to specify during the analysis phase.

We can distinguish three key moments for the evaluation, depending on the objectives of the project:

- the diagnostic evaluation;
- the formative evaluation;
- the summative evaluation.

The diagnostic evaluation takes place at the start of the process. It makes it possible to make an inventory of an existing artefact. For example, if a client already has a contract and encounters interaction concerns, it will involve establishing a diagnosis of the existing document. The objective is to determine the strengths and weaknesses of the current process in order to start on an existing basis. It is an inventory of what already exists, what could be improved, and understanding the reasons underlying potential issues. The measures of the diagnostic assessment can also be used as measure of improvement after redesign.

The formative evaluation comes later in the process, when the concepts have been defined. For example, if you need to design a new contract, you may have developed several concepts that correspond to different assumptions. In one concept, you might want to test the understandability of a metaphor, in the other the presence of graphics in addition to text. At this time, it is important to compare the effectiveness of the different solutions designed. Thus, formative evaluation is used during development to test concepts iteratively in order to consolidate them with feedback from future users.

Finally, summative evaluation generally takes place at the end of the process. At the end of a project, once the various iterative steps have been carried out, a finalized version of the artefact is tested in order to measure different parameters. The objective here is to know if the initial requirements and constraints have been met. This is the stage where the quality of the documents produced is evaluated in order to know whether the KPIs have been achieved.

11.3 WHAT TO EVALUATE?

The starting point is that documents are artefacts, just like any other objects, with which users are going to interact.

The quality of a documents is going to directly impact the quality of the interaction. In turn, the quality of the interaction will impact the action (or the absence of action) triggered by the document.

Obviously, the quality of a document is multi-factorial: it's a combination of acceptability/acceptation and usability.

Based on the existing literature in interaction design, we developed an evaluation framework matching contract design concerns and requirements:

Dimension	Sub-dimension	Definition
Acceptability / acceptance	Attitude	The **relationship of the user to the tasks to be performed and to the document.** Emotions, role of past experiences, confidence, engagment
	Perceived utility	The **degree to which the user thinks the document is useful and brings** something to him.
	Perceived usability	The degree to which the user thinks using **the artefact is easy and pleasant**
Usability	Effectivness	The document allows Accessibility, ease of learning, error handling
	Efficiency	The document allows the user to **complete the correct action by minimizing the necessary resources** Readability, mental load, comprehensibility, memorability / retention of information
	Satisfaction	The **user's feeling in his overall experience** with the document Hedonic qualities, emotions, attractivness, pleasure of use

Table 11.1 Amurabi's framework to assess legal documents in the User Testing Lab

Two main dimensions characterize the interaction with a legal document: acceptability and usability. Each dimension can be sub divided into sub-dimensions. The framework and the underlying models are described in the following sections.

11.3.1 Acceptability and Acceptance

'Intention to use' a technology, a service, a system, or even a document is based on several factors. These factors were widely studied in social sciences for decades. User intention to use a legal document is a key point for contract design because it directly affects user's action toward the document.

Here, the key point is to evaluate the 'degree of integration and ownership in a usage context'.[10]

Some authors distinguish between acceptability and acceptance, by specifying that - acceptability corresponds to the degree of usage intent in society (rather before the use of the artefact), while acceptance would depend from one individual to the other, and would be linked to one interaction in particular (during and after the use of the artefact).[11] Other authors use these terms interchangeably.

What we're trying to evaluate through acceptance and acceptability criteria is the degree of motivation of the user to interact with the artefact.

---

[10] Javier Barcenilla and Joseph Maurice Christian Bastien, 'Acceptability of New Technologies: Relationships with Ergonomy, Usability and User Experience?' (2009) 72(4) Le travail humain 311 [https://www.cairn-int.info/journal-le-travail-humain-2009-4-page-311.htm] accessed 23 April 2021.

[11] Jens Schade and Bernhard Schlag (eds), 'Acceptability of Transport Pricing Strategies: An Introduction' (Emerald Group Publishing Limited 2003) [https://www.emeraldinsight.com/doi/pdfplus/10.1108/9781786359506-001] accessed 23 April 2021.

The user's motivation will be key in its intent to use the artefact, and how it is going to use it. For example, users usually expect documents to meet their information needs. Their acceptance will thus depend upon the ability of the document to meet this need: this is the perceived usefulness criterion.

There are numerous acceptability models. The following models are widely acknowledged in the human sciences. They evolved over times and often complement each other. Every models aims at explaining why users actually use a system and what are the underlying factors.

11.3.1.1 Technology Acceptance Model (TAM)[12]

In this model, the perceived usefulness is introduced as a key notion. For example, a technology that is easy to use will not be used if it is perceived as useless. According to Davis, the usefulness plays a role twice as important as the perceived ease of use.

This is a historical model which was used as a framework for subsequent models.

11.3.1.2 Theory of planned behavior[13]

As for the TAM model, the usage intent would be influenced by the behavior and by beliefs on some factors influencing performance and control. These factors introduce a social component which may impact individual acceptance.

The introduction of social factors is important in the history of these models, as it is still valid and acknowledged today. This model is a landmark in the evolution, and has been enriched by subsequent models.

11.3.1.3 The Unified Theory of Acceptance and Use of Technology (UTAUT)[14]

These 8 models are based on the use of software. They enriched the previous ones through the presence of 'moderators' which influence the weight of the various dimensions (for example, usefulness would be more important for female users). This model is still considered valid and regularly updated.

11.3.1.4 The acceptability model[15]

This model distinguishes between acceptance before usage (mostly influenced by societal opinions), and acceptance after usage (called in practice acceptability). The perceived usefulness and the usability predict acceptance.

This model is commonly used and recognized, it is completed by the subdimensions of usability.

---

[12] Fred D Davis, 'Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology' (1989) 13 MIS Quarterly 319.

[13] Icek Ajzen, 'The Theory of Planned Behavior' (1991) 50 Organizational Behavior and Human Decision Processes 179.

[14] Viswanath Venkatesh and others, 'User Acceptance of Information Technology: Toward a Unified View' (2003) 27 MIS Quarterly 425.

[15] Jakob Nielsen, *Usability Engineering* (Morgan Kaufmann 1993).

11.3.1.5 Temporality of acceptance[16]

These authors point out that 48% of products which are brought back to the shops are functioning properly. This introduces the very important notion of dynamic and temporality in the adoption.

The previous models describe the factors which influence adoption, but these factors may be more or less significant during the actually lived experience. Karapanos describes 3 steps: orientation, incorporation and identification. Initially, acceptance would be based solely on pragmatic qualities; later on, the social and hedonist aspects would gain importance.

11.3.1.6 What do these various models tell us and how to apply them to contracts?

These models mostly tell us that usage is influenced by usage intent. Intent would be mainly linked to:

- Characteristics of given individuals
- Societal influences
- Usability and usefulness

It's also important to bear in mind that acceptance evolves over time and that so does the weight of the influencing factors.

How to apply these models to contracts and other legal documents?

Once familiarized with the various models, one should decide which heuristic criteria are the most important to be tested in a given project, depending on the purpose of the document, its context of usage, any reflex of the users to blind-sign or just ignore it.

In a project for a French public entity related to personal data protection, our mission namely consisted in creating interfaces which would help under-age users to better understand and exercise their rights.

Further to 'state-of-the-art' research, a thorough benchmark, several focus groups and co-creation workshops with children aged from 8 to 17 years old (divided into 3 age groups: 8–10, 11–14 and 15–17 years old), we developed a dozen of prototypes, which we tested both face-to-face with users (in a classroom), and during online conference calls.

The purpose of the interfaces was to:

- **Inform** underage users of their rights as regards their personal data;
- Collect valid **consent** from these users[17] within the meaning of the GDPR, ie 'informed consent';
- Facilitate these users' **exercise of their rights** on their personal data.

The usage context had been thoroughly analyzed through research, and confirmed by focus groups: underage users take for granted that they 'master interfaces' and think they know what to do with their personal data. In reality, the level of confusion and lack of understanding of their rights is quite high.

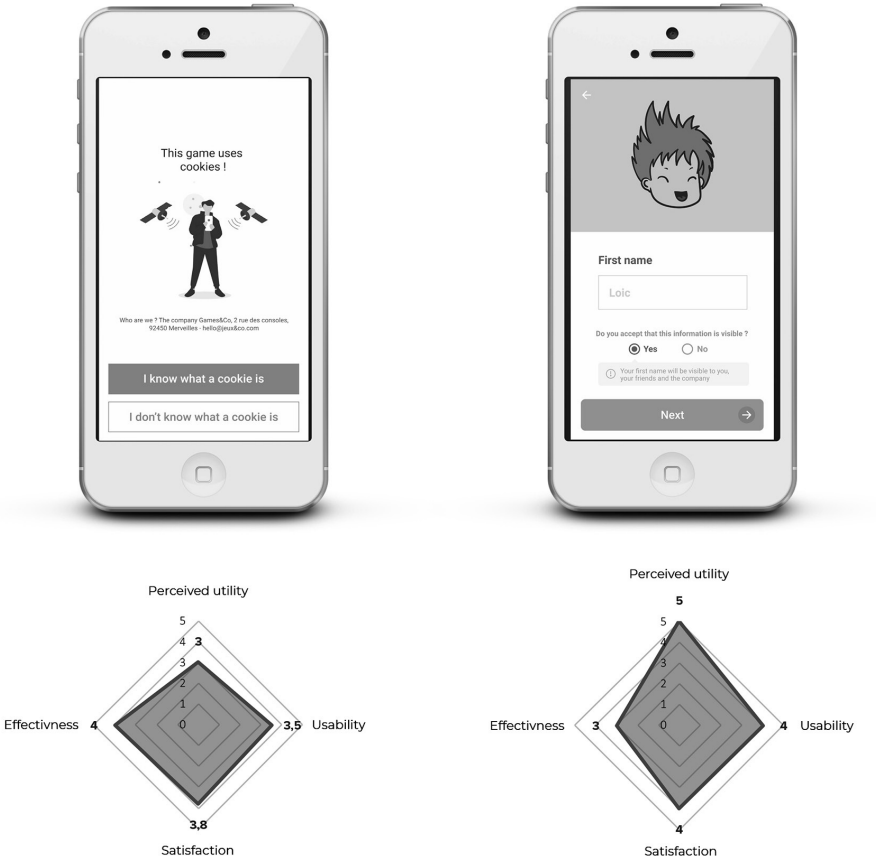
---

[16] Evangelos Karapanos and others, 'User Experience Over Time: An Initial Framework' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009)* (Association for Computing Machinery 2009).

[17] Users aged 15 and beyond may consent on their own, without their parents.

One of the key factors in the usage context was time, thus length of the text relating to the 'legal content'.

The test protocol included many different questions, and several were designed to measure acceptability:



Note: Prototype A3 was selected in the top three of deliverables. It was further improved based on both expert audits and user testing. The final version will be available under a Creative Commons license.

Figure 11.1 Privacy prototypes and testing for underage users

- Users' reactions before and after having seen the artefact,
- The perceived usefulness,
- The usage intent,
- A direct question: 'Would you like all other interfaces to have similar features?'

The results of the test clearly confirmed the overarching importance of the time required to perform the task and the length of the text: more than a couple of words is perceived as 'too much – not worth reading' by this age group.

Prototype A3 was selected in the top 3 of deliverables. It was further improved based on both expert audits and user testing. The final version will be available under a Creative Commons licence.

11.3.2 Usability

The ISO 9241-11 standard on usability describes it as:

> 'The extent to which a product can be used by specified users to achieve specified goals, with effectiveness, efficiency and satisfaction in a specified context of use'.[18]

Let's have a look at the various components of this definition:

Efficiency

- The user succeeds in doing what he had to do
- The performance of the completion of the task (goals are accurately reached)
- Ease of learning

Effectiveness

- The user meets his/her goals quickly and easily
- The completion of the task requires minimal resources

Satisfaction

- The user enjoys using the artefact.

Interestingly, the creator of Trello, Joel Spolsky, sees a connection between usability and 'a bit of human rights' and 'respect for humanity':

> Usability, fundamentally, is a matter of bringing a bit of human rights into the world of computer-human interaction. It's a way to let our ideals shine through in our software, no matter how mundane the software is. You may think that you're stuck in a boring, drab IT department making mind-numbing inventory software that only five lonely people will ever use. But you have daily opportunities to show respect for humanity even with the most mundane software.[19]

Are there more precise criteria to assess usability?

---

[18] ISO 9241, 'Ergonomics of Human System Interaction – Part 210: Human-Centered Design for Interactive Systems' (2010).

[19] Andreas Komninos, 'An Introduction to Usability' [https://www.interaction-design.org/literature/article/an-introduction-to-usability] accessed 23 April 2021.

11.3.2.1 Nielsen's five dimensions of usability

Indeed, Nielsen[20] provides 5 key dimensions of usability: ease of learning, efficiency, memorability, tolerance to errors and satisfaction.

11.3.2.2 The Hassenzahl UX model[21]

Hassenzahl distinguishes between the pragmatic (concrete and objective qualities, such as efficiency) and the hedonic qualities of the artefact (abstract and subjective qualities such as aesthetics, emotions generated by an artefact, attraction to it).

In this model, which is among the most recognized ones, the emphasis is put on the complementarity of the emotional aspects of the users (and the emotional curve, as a consequence), and the emotional aspects of the products (the emotional expression as hedonic attributes).

This model is particularly interesting in compliance design: one could create two anti-corruption guides which would be equally easy to read, swift to interact with, etc, but the hedonic qualities of the documents will play a key role in the acceptance and usability of the documents: if users consider one document as 'beautiful', reflecting the values of the company and generating a sense of pride, they will better interact with this document, leading to better implementation.

Additional criteria are also mentioned to describe usability:

- Accessibility is often mentioned. It describes the adequation of the document with the personal characteristics of the user (job, age, previous knowledge of the content, sensorial or physical abilities).
- Efficiency includes key criteria such as global and in-depth understanding, and memorization.
- Satisfaction includes the general quality of the user experience, taking into account the hedonic qualities of the artefact (eg aesthetics, attractivity).

11.4 HOW TO EVALUATE? EXPERT AUDIT AND USER TESTING

There are two main evaluation methods, which complement each other: the expert audit and the user-based evaluation. This chapter will provide both the evaluation methodology and concrete examples of implementation of the methodology to real documents, along with the results.

Clearly, these two methods of evaluation are totally complementary.

11.4.1 Expert Audit

Expert-based evaluations / Usability inspection methods.

---

[20] Nielsen, *Usability Engineering* (n 15) 26.

[21] Marc Hassenzahl, 'The Thing and I: Understanding the Relationship between User and Product' in Mark A Blythe and others (eds), *Funology: From Usability to Enjoyment* (Kluwer Academic Publishers 2003).

The expert audit is often used to complement user testing because this method is quite cost efficient. This type of evaluation is performed without the users, ie desk-based research. In a user-centric approach, the experts put themselves in the shoes of several typologies of users and in the context of several use cases.

Experts base their judgement on their previous experience (eg difficulties already met with other users), state of the art and guidelines.

In order to reduce individual biases, it is useful to combine several experts for a given evaluation.

Let's be clear; expert evaluation will never replace user testing, but it is useful to:

- Quickly formulate recommendations
- Early detect the main defects

To avoid subjectivity and personal bias, experts analyze the compliance of the artefact with a set of guidelines, standards, and ergonomic criteria, also called ' heuristics '. Heuristics provide an objective framework to work with.

There are numerous heuristic evaluation grids, based on fundamental ergonomic principles. However, it's preferable to create one's own analysis grid, based on the criteria which are the most relevant for the artefact and users concerned.

Which are the main heuristic criteria to evaluate the quality of a document?

11.4.1.1 ISO Norms

ISO 9241-110:2020[22] provides a list of 7 heuristic principles for interaction between a user and a system:

- **Suitability for the user's task**: the system enables the user to perform the task efficiently and effectively
- **Self-descriptiveness**: the system is understandable at first sight thanks to the system feedback or can be explained at the user's request
- **Controllability**: the user can initiate and control the direction and the rhythm of the interaction
- **Conformity with users' expectations**: the system is consistent and corresponds to the characteristics of the users
- **Use error robustness**: the goal can be achieved with or without corrective actions
- **Capacity to individualization**: the interface can be modified to adapt to the needs of the task or to users' preferences or skills
- **Learnability**: the dialogue supports and guides the user in learning about the system.

In addition, ISO 9241-112[23] also provides a list of heuristics for the presentation of information:

- **Detectability**: information is provided at the right time, at the right place
- **Clarity**: the content is displayed quickly and precisely
- **Discriminability**: the various information can be distinguished precisely from one another

---

[22] ISO 9241-110:2020, 'Ergonomics of Human-System Interaction — Part 110: Interaction Principles' (2020).

[23] ISO 9241-112:2017, 'Ergonomics of Human-System Interaction — Part 112: Principles for the Presentation of Information'.

- **Legibility**: the information is easy to read
- **Unambiguous interpretability**: the meaning of the terms is clearly understandable
- **Conciseness**: only the information necessary to perform the task are displayed
- **Consistency**: similar information are displayed in a similar way throughout the interface

11.4.1.2 Scapin & Bastien

In addition, Scapin & Bastien ergonomic criteria for human-computer interfaces (1997)[24] remain key and state-of-the-art.

Scapin & Bastien's work widely relate to how to incorporate human factors considerations into the process of designing and evaluating human-computer interfaces.

Based on a state-of-the-art of about 800 studies, these authors were able to identify a list of 8 criteria. These heuristics are well recognized and widely used for expert evaluations.

Guidance

This criterion refers to all the means available to advise, orient, inform the user (eg labels, alarms, messages), including the language used. Interestingly, the authors' rationale for this criterion is that good guidance enables the user to better learn how to use a given system. There's an obvious connection with the empowerment principle in contract design where one should always draft and design to autonomize the user.

Workload

This criterion concerns all interface elements which help reduce the user's perceptual or cognitive load, and help increase the efficiency of the dialogue. Obviously, the rationale is that limiting the workload limits the risk of errors.

Explicit Control

The relationship between the functioning of the interface and the actions of the the user must be explained very clearly. In addition, the user should always be in control of the system processing (eg interrupt, pause, continue…).

Adaptability

The interfaces should be able to adapt to various users' operating modes. The system must respect the level of experience of the user, eg by allowing experienced users to bypass a preliminary step.

Error management

Setting up means to detect and prevent errors, ensure that the information provided to user sis clear as to the nature of the error made (language, format), and ensure that the actions to correct errors are relevant and easy to read. Provide users with easy means to correct errors.

Consistency

The choices made to design the interface (naming, formats, colors…) must be similar for similar contexts, and different for different contexts.

---

[24] Dominique L Scapin and JM Christian Bastien, 'Ergonomic Criteria for Evaluating the Ergonomic Quality of Interactive Systems' (1997) 16 Behaviour & Information Technology 220.

Significance of codes
This qualifies the relationship between a term or a symbol and its reference. There should be no ambiguity at all, particularly with regards to icons. Words and icons chosen must respect conventions.

Compatibility
The interface must adapt to the users' operating mode, and respect established standards of accessibility.

One could also rely on Nielsen's 10 heuristic principles[25], and/or conformity with human factors guidelines, with the following checklist:

- Number of fonts
- Font size
- Styles
- Function of colors
- Contrasts
- Content and paragraphs
- Hierarchy of the information, etc.

In contract design, accessibility of the content to the widest number of users is key. Thus, one could also rely on the Web Content Accessibility Guidelines (WCAG), which provides precise criteria for the wide accessibility. Fast testing tools are also available. For example, the Stark plug-in visualizes what color-blind users see, and measures contrasts.

11.4.1.3 How might we evaluate legibility … without readers?

Many tests and formulae exist to to provide a score of ease of reading, or level of education required to understand a text upon first reading: legibility formulae of Dale-Chall[26], Simple Measure of Gobbledygook SMOG[27], etc).

However, these tests are mostly interesting in relative evaluations (assessing one option against another), and must always be completed by user testing.

For example, in Flesch legibility formulae, the score is calculated as:

206.835 – (1.015 x ASL) – (84.6 x ASW) with:

- ASL = average length of sentence (number of words divided by the number of sentences)
- ASW = average number of syllabus per word (number of syllabus divided by the number of words)

---

[25] Jakob Nielsen, 'How to Conduct a Heuristic Evaluation' (Nielsen Norman Group 1995) 1, 1–8.

[26] Edgar Dale and Jeanne S Chall, 'A Formula for Predicting Readability' (1948) 27 Educational Research Bulletin 11.

[27] G Harry McLaughlin, 'SMOG Grading: A New Readability Formula' (1969) 12 Journal of Reading 639.

The text is evaluated on a scale of 100 points. The higher the score, the easier it will be to understand the document upon first reading. For standard texts, a satisfactory score would be between 60 and 70.

These tests are interesting as a reference, but they completely overlook the usage context, user profiles and journeys so they should always be used in combination with other quality measures of the acceptability and usability of legal documents.

Based on all these standards and research, we developed for example the following analysis grid for expert audit, to be adapted depending on the purpose of the document, and its usage context:

Dimension	Sub-dimension	Measures
Acceptability / acceptance	Attitude	**Data analytics** (e.g. opening rates, participation rates, number of incidents, support requests) **Partners feedbacks**
	Perceived utility	
	Perceived usability	
Usability	Effectivness	**Data analytics (e.g. nature of incidents, support content)** Accesibility : **WCAG** criteria Potential interaction problems detected on the basis of analysis grid (nature, extent, severity) **Flesch readibility index (score from 0 to 100)**
	Efficiency	
	Satisfaction	

Table 11.2 Amurabi's framework for expert audit in the User Testing Lab

Understandability of contracts is obviously a key concern in legal innovation by design projects. Plain language methodology is often perceived by lawyers as 'law for dummies', and a loss of expertise on their side. In reality, it's quite the opposite, it takes the very best lawyers and contract experts to make them accessible to users.

In order to overcome this resistance to change, we've developed a simple and visual way of raising awareness among lawyers as to the complexity of the words and sentences they use.

We apply color codes to the document at stake, clearly visualizing which words or expressions are:

- Complex or formal (they appear below in yellow)
- Jargon (in red)
- Negative form, which is disengaging (light grey)
- Passive form, which is disengaging and more difficult to understand (dark grey)
- Theoretical or imprecise (kaki)

11.4.2 User-based Evaluations

User testing, questionnaires, interviews, focus groups are part of evaluation methods requiring the direct participation of users.

User scan either perform representative tasks for which the document has been designed, or freely explore the document.

The goal of these tests is to identify usage difficulties, based on:

- Verbalizations from the users, either induced or spontaneous
- Performance indicators such as the time required to complete a given task, the number and the type of errors made, etc.

There are numerous ways to test artefacts with users. We briefly present below the methods which are relevant for legal documents, and fairly easy to implement.

11.4.2.1 Think aloud

In this test methodology, the user interacts with the artefact to perform a specific task, or during a free interaction. During the interaction, users are being asked to voice their ideas, beliefs, expectations, doubts, discoveries, etc.

This test is often used when usage data are recorded, in order to explain the user journey on an interface.

Verbalizations can be simultaneous (during the interaction), or retrospective (once the task has been completed, with or without the possibility to see a video of the various actions made).

Simultaneous verbalizations are generally preferred as they eliminate the possibility that users be selective in the way they recall or rationalize their actions.

For example, in the data privacy project for underage users, we created small usage scenarios of the various prototypes, and asked users to comment out loud their actions, where they click and why.

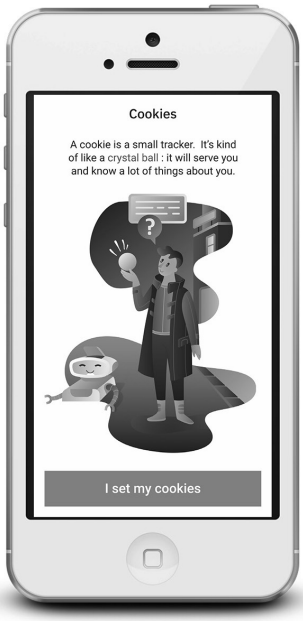


Figure 11.2 Data privacy prototype for underage users

This prototype obtained a low score in understandability, because the crystal ball was confusing. The key verbatim was: 'This is not clear enough, because of the crystal ball. I don't understand why they tell us about a crystal ball at the same as cookies.'

This prototype was abandoned.

11.4.2.2 Questionnaires and interviews

With questionnaires, the evaluation is indirect. Questionnaires are subjective evaluation tools, given that they reflect the users' opinion. They are generally filled-in by the user when he/she is alone. That's why it's important to generate confidence upfront, for example by inserting a welcome message:

> 'Please answer intuitively, according to your feelings when using the artefact. There is no right or wrong answer: only your opinion matters.'

Before building the questionnaire, it is necessary to have a clear idea of the objective: validate a hypothesis, test a specific feature of the artefact, or evaluate a concept.

For example, in a project consisting in designing a compliance code of conduct for a large group (which had to be accepted and signed by employees as part of their employment agreement), the main purpose of the document was to (i) trigger reading instead of pushback and (ii) ensure in-depth understanding of fairly complex notions.

Among the questions asked, we inserted this question, as a proxy for reading motivation which was biased because of the test itself. For example, 'What is the first word that came to your mind when you opened the document?'

We also inserted a range of questions, either basic, multiple choice or open, more complex questions to test understanding (attributing specific grades to each right answer, and comparing 2 groups of users, one with the original document, and the other with the redesigned one). For example, 'How much of the content did you understand, on a scale from 1 (lowest) to 10 (highest)?', followed by control questions in the form of short scenarios, eg 'Gifts are customary in Russia. I am meeting a government executive on a business trip to Saint Petersburg. Can I offer him a range of our products? Yes / No / Explain why'.

It is possible to rely on normalized questionnaires, which have been scientifically tested, using for example the Rating Mental Scale Effort[28], Van der Laan Acceptance Scale[29], the User Experience Questionnaire[30], the AttrakDiff scale (short-version)[31] or the System Usability

---

[28] Fred R H Zijlstra and Leendert van Doorn, 'The Construction of a Subjective Effort Scale' (Report, Department of Social Sciences and Philosophy, Delft University of Technology 1985).

[29] Jinke D Van Der Laan, Adriaan Heino and Dick De Waard, 'A Simple Procedure for the Assessment of Acceptance of Advanced Transport Telematics' (1997) 5 Transportation Research Part C Emerging Technologies 1.

[30] Bettinna Laugwitz, Theo Held and Martin Schrepp, 'Construction and Evaluation of a User Experience Questionnaire' in Andreas Holzinger (ed), *HCI and Usability for Education and Work. 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20–21, 2008. Proceedings* (Springer 2008).

[31] Marc Hassenzahl, Michael Burmester and Franz Koller, 'AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität' in Jurgen Ziegler and Gerd Szwillus (eds), *Mensch & Computer 2003. Interaktion in Bewegung* (BG Teubner 2003).

Scale[32]. It is also perfectly valid to create your own questionnaire, if you pay attention to the biases.

- **The desirability bias**: the user may answer depending on what he thinks the tester's expectations are. To avoid this bias, the introductory message is important, as well as anonymizing the answers and letting the user fill-in the questionnaire without looking at what he/she does if the test is performed in a meeting room for example;
- **The halo effect**: spread the questions in various parts of the questionnaire when the answer to one might strongly influence the answer to others;
- **The suggestion bias**: the way the question is drafted should not suggest a given answer;
- **The defensive contraction bias**: personalized questions (such as 'do you think that…?') do not always favor self expression. Quite the opposite, such wording may generate escaping answers;
- **Positive answer attraction** (tendency to concur): to avoid this bias, you may use reverse formulations or confirm a given result with a second question.

Interviews offer greater flexibility and enables the testers to quickly react and build upon users' answers.

11.4.2.3 Collaborative evaluation during workshops

Evaluations during workshops enable to combine activities designed to get individual feedback, and collective activities to reach a consensus for example on how to improve a given feature.

In these workshops, we often start with a usage scenario and a task to be completed within a given timeframe, with precise grades corresponding to the number of right answers within the time limit.

For example, for the redesign of a B-to-C data protection policy, the test workshops included some basic questions, like 'do you feel like clicking?'

And usage scenarios, such as 'tell us who can see and use your personal data?' after having displayed the relevant part of the prototype.

Then, you may conduct more recreational activities such as the Cloze Test, or the 5 Seconds Test (see below).

Before ending the workshop, it is useful to present the results of the tests to the users, and collectively explore solutions to the identified problems.

The 'cloze procedure' results from Wilson L Taylor's research.[33] 'Closure' refers to a concept known in the Gestalt psychology, which Taylor applies to reading:

> 'The mind fills the gaps!'

In practice, one presents to the users a text in which one has removed (randomly or regularly) one word every N word of the text (10 to 20% of suppressed words, often N=6).

---

[32] John Brooke, 'SUS: A Quick and Dirty Usability Scale' (November 1995) [https://www.researchgate.net/publication/228593520_SUS_A_quick_and_dirty_usability_scale] accessed 10 August 2021.

[33] Wilson L Taylor, '"Cloze Procedure": A New Tool for Measuring Readability' (1953) 30 Journalism Quarterly 415.

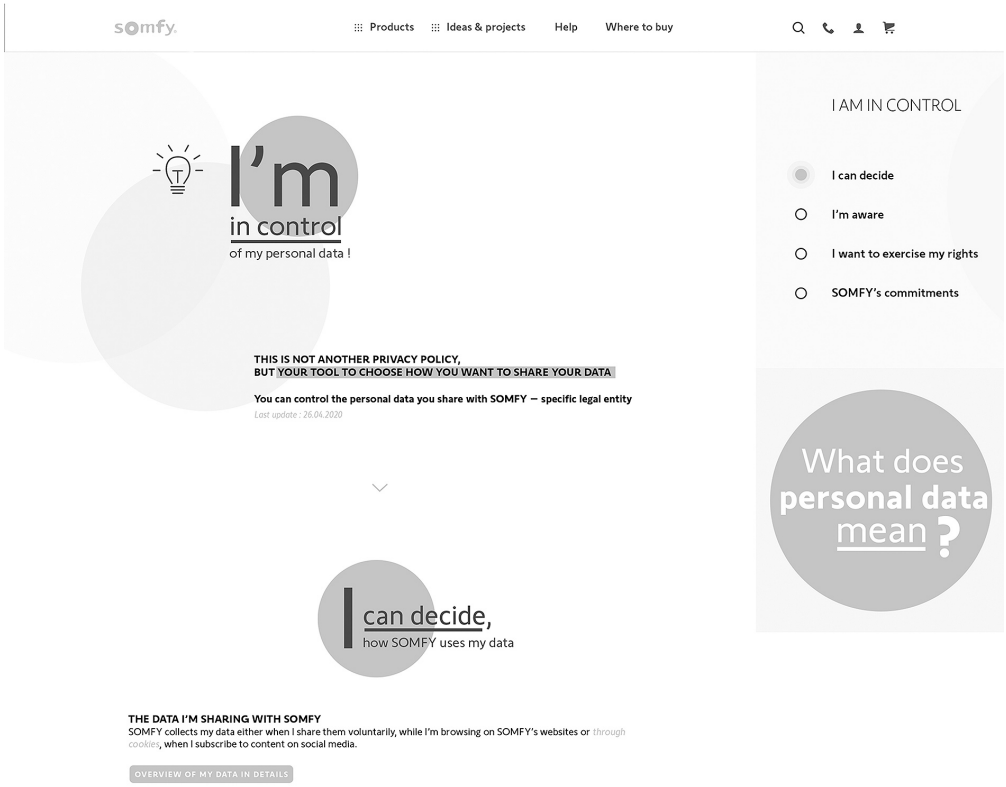


Figure 11.3 Data privacy charter

The user must fill-in the blanks. The percentage of right answers is calculated. 60% of right answers is considered as a satisfactory score.[34]

In this test, the redesigned document obtained 91% of right answers.

According to Lindgaard and others[35], the first impression is forged in 50 milliseconds, from graphic and visual elements. The first impression impacts acceptability (for example the perceived credibility of a given website, the perceived ease of use…).

In practice, before presenting the prototype, one gives to the users the following instruction: 'We are going to present a page during 5 seconds. Afterwards, we will ask you to tell us what you have seen.'

After having displayed the artefact during 5 seconds, users are asked to say what he/she remembers, but also his/her general impression (from 0 to 5), the perceived goals of the artefact, its aesthetics, etc.

---

[34] See for example Jakob Nielsen, 'Close Test for Reading Comprehension' (*Nielsen Norman Group*, 28 February 2011) [https://www.nngroup.com/articles/cloze-test-reading-comprehension/] accessed 23 April 2021.

[35] Gitte Lindgaard and others, 'Attention Web Designers: You Have 50 Milliseconds to Make a Good First Impression!' (2006) 25 Behaviour & Information Technology 115.

11.4.2.4 Online testing

Online questionnaires may be used to reach a large number of users. Several variations of the artefact may be tested, with different groups of users. In this case, the variations are attributed to the users in a random way. If possible, one should ensure that the user groups are consistent in terms of demography (location, gender, age…).

The questions are asked are always the same for all the variations of the artefact (understandability, intention of use, aesthetics etc.). Statistical tools can help to compare results if they significantly differ according to the variations of the artefact.

We developed for example the following analysis grid, to be adapted depending on the purpose of the document and its usage context:

Dimension	Sub-dimension	Objective measure	Subjective measure
Acceptability / acceptance	Attitude		
	Perceived utility		Van der Laan Acceptance Scale (perceived usefulness)
	Perceived usability		Van der Laan Acceptance Scale (perceived ease of use)
Usability	Effectivness	Number of erros % of goals achieved Number of errors after a long period of non-use	Nature of errors Questionnaires Verbalizations Sus questionnaires (System Usability Scale) UX measures (e.g. Attrakdiff)
	Efficiency	Reading speed Tasks completion time Number of tasks completed in a predefined time % of correct answers ito comprehension questions Rating Mental Scale Effort Score Performance test - retest after 3 months / 6 months (e.g. number of errors after a long period of non-use	
	Satisfaction		

Table 11.3 Amurabi's framework for user testing in the User Testing Lab

11.4.2.5 Iteration and long-term evaluation

Obviously, legal innovation by design is an iterative process. Further to testing, sufficient time must be allowed to the resolution of the problems identified during the test, combining the results of the expert audit and the user testing.

Before completion of the project, it is possible to test the redesigned version in Beta testing. The final prototype is put in the real environment of the client ('real world testing'). This enables the measurement of the impact of external factors, and improve the prototype further.

After completion, the project is actually not over. This is the time to follow up the key performance indicators you have defined at the beginning of the project, during 6 to 12 months. Do not underestimate the time and logistics required for this phase: follow-up with the client, setting-up some distance tracking tools (logs, analytics, etc.).

Do not forget simple indicators, such as lead time. For example, a redesigned NDA enabled a pharmaceutical laboratory to divide its lead time by 2.[36]

---

[36] See Amurabi, 'Be NDA-Ready and Compliant in No Time' [https://www.amurabi.eu/en/projects/be-nda-ready-and-compliant-in-no-time/] accessed 23 April 2021.

11.5 CONCLUSION

The scientific and data driven approach for testing the efficiency, usability, legibility of contracts, compliance programs, data privacy charters etc, provides a useful framework for legal innovators on a day-to-day basis: measuring the value might help to get internal sponsors, or even budget for the next project. It also creates a sense of pride among lawyers taking part of the project, and a great dynamics within legal divisions or law firms.

Limitations and further research: user testing and even more so expert audit may be limited in scope. Apart from online testing, user testing is usually limited to a few dozens of users for timing or budget reasons. While we've seen that there are ways to ensure representativity, the limited number of users involved calls for (i) care when analyzing the results, (ii) ideally, other user testing in different jurisdictions, which could be undertaken through some sort of open-source material, and (iii) continuous improvement in our ever-evolving environment.

In addition, measuring the value of contract design projects is not an end in itself. It's a way to leverage knowledge from one project to the next, and thus in turn reach systemic impact.

Further research and practice are needed to find ways to put in common learnings while ensuring confidentiality of sensitive information when need be.

BIBLIOGRAPHY

Ajzen I, 'The Theory of Planned Behavior' (1991) 50 Organizational Behavior and Human Decision Processes 179.

Amurabi, 'Be NDA-Ready and Compliant in No Time' [https://www.amurabi.eu/en/projects/be-nda-ready-and-compliant-in-no-time/] accessed 23 April 2021.

Barcenilla J, Maurice J and Bastien C, 'Acceptability of New Technologies: Relationships with Ergonomy, Usability and User Experience?' (2009) 72(4) Le travail humain 311 [https://www.cairn-int.info/journal-le-travail-humain-2009-4-page-311.htm] accessed 23 April 2021.

Belley JG, 'Max Weber et la Théorie du Droit des Contrats' ('Max Weber and the Theory of Contract Law') (1988) 9 Droit et Société (Law and Society) 281.

Brooke J, 'SUS: A Quick and Dirty Usability Scale' (November 1995) [https://www.researchgate.net/publication/228593520_SUS_A_quick_and_dirty_usability_scale] accessed 10 August 2021.

Dale E and Chall JS, 'A Formula for Predicting Readability' (1948) 27 Educational Research Bulletin 11.

Davis FD, 'Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology' (1989) 13 MIS Quarterly 319.

Hassenzahl M, 'The Thing and I: Understanding the Relationship between User and Product' in Blythe MA, Monk AF, Overbeeke K and Wright PC (eds), *Funology: From Usability to Enjoyment* (Kluwer Academic Publishers 2003).

Hassenzahl M, Burmester M and Koller F, 'AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität' in Ziegler J and Szwillus G (eds), *Mensch & Computer 2003. Interaktion in Bewegung* (BG Teubner 2003).

Karapanos E, Zimmerman J, Forlizzi J and Martens JB, 'User Experience Over Time: An Initial Framework' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009)* (Association for Computing Machinery 2009).

Komninos A, 'An Introduction To Usability' [https://www.interaction-design.org/literature/article/an-introduction-to-usability] accessed 23 April 2021.

IACCM, 'The Purpose of a Contract – Research Report' (2017).

Laugwitz B, Held T and Schrepp M, 'Construction and Evaluation of a User Experience Questionnaire' in Holzinger A (ed), *HCI and Usability for Education and Work. 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20–21, 2008. Proceedings* (Springer 2008).

Lindgaard G, Fernandes G, Dudek C and Brown J, 'Attention Web Designers: You Have 50 Milliseconds to Make a Good First Impression!' (2006) 25 Behaviour & Information Technology 115.
McLaughlin GH, 'SMOG Grading: A New Readability Formula' (1969) 12 Journal of Reading 639.
Nielsen J, *Usability Engineering* (Morgan Kaufmann 1993).
Nielsen J, 'How to Conduct a Heuristic Evaluation' (*Nielsen Norman Group* 1995).
Nielsen J, 'Close Test for Reading Comprehension' (*Nielsen Norman Group*, 28 February 2011) [https://www.nngroup.com/articles/cloze-test-reading-comprehension/] accessed 23 April 2021.
Passera S, 'Flowcharts, Swimlanes, and Timelines: Alternatives to Prose in Communicating Legal–Bureaucratic Instructions to Civil Servants' (2018) 32 Journal of Business and Technical Communication 229.
Scapin DL and Bastien JMC, 'Ergonomic Criteria for Evaluating the Ergonomic Quality of Interactive Systems' (1997) 16 Behaviour & Information Technology 220.
Schade J and Schlag B (eds), 'Acceptability of Transport Pricing Strategies: An Introduction' (Emerald Group Publishing Limited 2003) [https://www.emeraldinsight.com/doi/pdfplus/10.1108/9781786359506-001] accessed 23 April 2021.
Taylor WL, '"Cloze Procedure": A New Tool for Measuring Readability' (1953) 30 Journalism Quarterly 415.
Van Der Laan JD, Heino A and De Waard D, 'A Simple Procedure for the Assessment of Acceptance of Advanced Transport Telematics' (1997) 5 Transportation Research Part C Emerging Technologies 1.
Venkatesh V, Morris MG, Davis GB and Davis FD, 'User Acceptance of Information Technology: Toward a Unified View' (2003) 27 MIS Quarterly 425.
Zijlstra FRH and van Doorn L, 'The Construction of a Subjective Effort Scale (Report, Department of Social Sciences and Philosophy, Delft University of Technology 1985).