An Empirical Investigation of Software Engineers' Ability to Classify Legal Cross-References

Jeremy C. Maxwell
Dept. of Computer Science
North Carolina State University
Raleigh, NC, USA
jcmaxwe3@ncsu.edu

Annie I. Antòn
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA, USA
aianton@cc.gatech.edu

Julie B. Earp
College of Management
North Carolina State University
Raleigh, NC, USA
julie_earp@ncsu.edu

***Abstract*—Requirements engineers often have to develop software for regulated domains. These regulations often contain cross-references to other laws. Cross-references can introduce exceptions or definitions, constrain existing requirements, or even conflict with other compliance requirements. To develop compliant software, requirements engineers must understand the impact these cross-references have on their software. In this paper, we present an empirical study in which we measure the ability of software practitioners to classify cross-references using our previously developed cross-reference taxonomy. We discover that software practitioners are not well equipped to understand the impact of cross-references on their software.**

***Index Terms*—Cross-References, Regulatory Compliance, Requirements, Software Engineering**

I. Introduction

Laws and regulations govern organizations and the software systems that they use. Requirements engineers must understand these laws and regulations as they are specifying software requirements to ensure regulatory compliance in software systems. Requirements engineers need tools and techniques to adapt to regulatory evolution. A regulatory text can change as often as once a year [1], requiring potentially critical modifications to software. As laws and regulations evolve over time, engineers must understand the impact on their systems and adapt their software accordingly. To this end, requirements engineers need to (a) understand the impact that legal cross-references have on software, and (b) be able to identify the conflicts that cross-references sometimes introduce.

In the United States, regulations are issued by federal agencies that regulate particular domains. For example, the Department of Health and Human Services (HHS) regulates healthcare-related industries. When a regulatory agency seeks to issue a new regulation, the agency will first issue a proposed rule or a notice of proposed rule making (NPRM). Except in emergencies, the public will then be given the opportunity to comment on the proposed rule. The regulatory agency then issues a final rule that is binding on the regulated domain. Oftentimes, regulatory deadlines are too compressed and do not allow engineers the luxury of waiting for final regulations to be published. For example, the American Recovery and Reinvestment Act of 2009 (ARRA) created the Meaningful Use (MU) program that makes $23 billion in incentives available for healthcare providers that adopt certified EHR technology and use it in a meaningful way. The incentives are paid out over three Stages that require providers to meet increasingly intensified clinical quality criteria. For example, one of the clinical quality criteria concerns patient engagement; for each stage of MU, providers must engage a greater portion of their patients and in more ways. As part of the criteria, EHR technology must be updated during each stage of MU to enable physicians to document, track, and submit the clinical quality criteria. The proposed rule for MU Stage 1 was released on January 13, 2010, whereas the final rule was issued on July 28, 2010. Eligible providers and hospitals could begin applying for Stage One incentives on January 1, 2011. Engineers that waited until the final rule was released were left with less than six months to adapt their EHRs to meet the MU Stage 1 requirements, have their EHR certified, and installed at physician practices and hospitals. EHR vendors have stated that these timeframes are too short [2]. The inability to comply with short compliance timeframes is not limited to EHR vendors. For example, the original GLBA Financial Privacy Rule proposed a compliance date six months after the final rule was published; commenters responded that this timeframe was too short to comply with the rule, and instead asked for a 12 to 24 month compliance timeframe.

Because market forces so often compel software organizations to be the first to market, they must begin complying with regulations before the final rule is published. This introduces significant ambiguity and uncertainty when building software that must comply with these changing legal requirements. Our prior work [3]–[6] seeks to provide engineers with tools and techniques to develop software within changing regulatory environments. In particular, we developed a taxonomy for classifying the impact that legal cross-references have on software requirements, particularly when they introduce conflicts. Misclassification of legal cross-references can lead to costly noncompliance. Herein, we present the findings of an empirical study in which we find that software practitioners are not well-equipped to understand the impact of compliance requirements on software.

The remainder of this paper is organized as follows: Section II outlines related work; Section III provides background on legal cross-references; Section IV presents an overview of our research design; Section V summarizes participant

978-1-4673-5765-4/13/$31.00 © 2013 IEEE

performance; Section VI discusses our results; Section VII outlines threats to validity; and Section VIII concludes this work.

II. Related Work

We discuss our prior work in Section III. In this section, we provide an overview of related work in requirements engineering and regulatory compliance. Prior researchers have focused on: eliciting compliance requirements using frames [7] and production rules [3]; evaluating business processes for compliance [8]; compliance across jurisdictions [9]; and traceability from regulations to software requirements [10].

Researchers have also empirically studied how requirements engineers approach the law [11]. Breaux assessed the ability of requirements engineering graduate students to acquire regulatory requirements using a legal ontology [11]. He found that students could not agree on which legal statements matched his ontological concepts unless provided with additional tools—in Breaux's case, phrase heuristics [11]. Massey et. al assess the ability of software engineering graduate students to identify software requirements that are legal implementation ready (LIR) and find that students are not well equipped to make these decisions [12]. Young et al. compare the effectiveness of goal-based approaches to commitment-based approaches for extracting compliance requirements from privacy policies, and discover that a multiple approaches yields the best results [13]. All of these studies, however, employ graduate students. To the best of our knowledge, we are the first researchers that analyze the ability of software practitioners to analyze compliance requirements.

III. Legal Cross-References

In this section, we provide an overview of legal cross-references and our cross-reference taxonomy. A legal cross-reference is a citation from one portion of a legal text to another portion of that text or to another text [4], [5]. The *referencing text* is the legal text that contains the cross-reference and the *referenced text* is the legal text that is cited. Cross-references establish relationships and priorities among laws at various levels. In our previous work studying healthcare and financial regulations, we encountered cross-references to statutory law (such as the U.S. Code), administrative law (such as the Code of Federal Regulations), executive orders issued by the President, as well as cross-references to a general set of laws [4], [5]. Cross-references can reference legal compilations by reference; for example, 22 U.S.C. 2709(a)(3) is a citation to Title 22, section 2709, subsection (a), paragraph (3) of the U.S. Code. Other times, cross-references may reference an act by its common name, for example, a reference to the Fair Credit Reporting Act. An additional area of law—case law—is created by judicial decisions in response to court cases. Case law provides further explanation and interpretation of statutes and regulations. We have not yet encountered references to case law in our studies [5].

When a legal text governs software systems, cross-references may introduce constraints, exceptions, or even conflicts [4]. We have developed a cross-reference taxonomy to help requirements engineers understand the impact that cross-references have on software [5]. Table I summarizes the classifications in our taxonomy [4], [5]. It is important to note that a given cross-reference may have multiple classifications. In the remainder of this section, we describe each classification in our taxonomy. For emphasis, we also italicize cross-references in the examples that follow.

TABLE I: Cross-Reference Taxonomy

Constraint
Exception
Definition
Unrelated
Incorrect
General
Prioritization

Constraint Cross-References Cross-references refine existing software requirements by introducing additional constraints are called constraint cross-references. Here is an example constraint cross-reference from the Gramm-Leach-Bliley (GLB) Financial Privacy Rule[1]:

> 16 CFR 313.6(a)(7): The initial, annual, and revised privacy notices that you must provide [...] must include each of the following items of information: [...] any disclosures that you make *under section 603(d)(2)(A)(iii) of the Fair Credit Reporting Act (15 U.S.C. 1681a(d)(2)(A)(iii)* (that is, notices regarding the ability to opt out of disclosures of information among affiliates)

Exception Cross-References Exception cross-references add exception conditions to software requirements. Here is an example exception cross-reference from the GLB Financial Privacy Rule:

> 16 CFR 313.15(a)(5)(i): The requirements for initial [privacy] notice [...] do not apply when you disclose nonpublic personal information to a consumer reporting agency *in accordance with the Fair Credit Reporting Act (15 U.S.C. 1681 et seq.)*

Definition Cross-References Legal texts use cross-references to cite definitions from other laws in much the same way as a programmer imports object and function definitions from language libraries. Here is an example definitional cross-reference in the GLBA Financial Privacy Rule:

> 16 CFR 313.3(f): Consumer reporting agency has the same meaning *as in section 603(f) of the Fair Credit Reporting Act (15 U.S.C. 1681a(f)).*

Unrelated Cross-References Unrelated cross-references are not relevant to software systems, because they do not introduce software requirements for the system. For example, the following cross-reference from the HIPAA Privacy Rule is not relevant to software systems, because the way an institutional research board is formed is beyond the scope of a software system.

[1] 16 CFR Part 313

> 45 CFR 164.512(i)(a)(i)(A): A covered entity may use or disclose PHI for research, regardless of the source of funding for that research, provided that the covered entity obtains documentation that an authorization or waiver, in whole or in part, of the individual authorization required by §164.508 for use or disclosure of PHI has been approved by either: an institutional review board (IRB), established *in accordance with 7 CFR 1c.107* [...]

Incorrect Cross-References Sometimes, updates to law make a cross-reference point to an incorrect portion of the law. For example, section 604(b)(3)(A)(ii) of the Fair Credit Reporting Act[2] contains a cross-reference to 609(c)(3) of the same Act.

> FCRA 604(b)(3)(A)(ii): Except as provided in subparagraph (B), in using a consumer report for employment purposes, before taking any adverse action based in whole or in part on the report, the person intending to take such adverse action shall provide to the consumer to whom the report relates—[...] a description in writing of the rights of the consumer under this title, *as prescribed by the Federal Trade Commission under section 609(c)(3)*

However, this reference no longer exists after Congress updated the Fair Credit Reporting Act. This fact is documented in a footnote in the legal text.

General Cross-References General cross-references are cross-references that do not reference a specific law. Here is an example general cross-reference in the GLBA Financial Privacy Rule:

> 16 CFR 313.17: This part shall not be construed as superseding, altering, or affecting *any statute, regulation, order, or interpretation in effect in any State* [...]

Often, general cross-references have multiple classifications. In the example above, the cross-reference is both a general and prioritization cross-reference.

Prioritization Cross-References Prioritization cross-references establish the priority between new law and existing law. Here is an example prioritization cross-reference in the GLBA Financial Privacy Rule:

> 16 CFR 313.16: Nothing in this part shall be construed to modify, limit, or supersede the operation *of the Fair Credit Reporting Act (15 U.S.C. 1681 et seq.)*

IV. Research Design

In this section, we describe the design of our user study. Our user study seeks to test the ability of software engineers, legal domain experts, and healthcare professionals to correctly classify cross-references using the cross-reference taxonomy we previously developed [4], [5]. We compare participant responses to classifications made by a group of experts including the author, a law school professor, and a software privacy professor. We measure precision to test participants' ability to make classifications. *Precision* measures consistency, or whether the classifications made by participants are the same classifications made by the experts. *Recall* is often measured with precision (e.g., [14]). Recall measures completeness, or whether participants identify all cross-references in the regulation. We do not measure recall in our user study. To measure recall, we could have provided participants with large paragraphs of legal text and asked them to identify cross-references in the text (to measure recall) then asked them to classify each (to measure precision). However, we determined that this exercise would have taken participants too much time and would have likely led to reduced participation by our target population (software and healthcare practitioners). Thus, we decided to ask participants to just do classification tasks and left identification tasks to future studies. To test precision, we use a set of cross-reference classifications that experts have identified and classified in healthcare and financial regulations.

Our study has the following null hypothesis:

> *H_1:* Individuals from the participant group have equivalent or greater precision than the expert classifications when classifying cross-references using the taxonomy.

A. Materials

The study materials consisted of an informed consent form, a demographics survey, a tutorial, and a 10 question survey. The demographics survey captured participant work experience, education, and comfort level with making security, privacy, and legal decisions. The tutorial explained our cross-reference taxonomy with examples of each classification. The survey consisted of 10 legal statements drawn from four healthcare and financial regulations:

- the Healthcare Portability and Accountability Act (HIPAA) Privacy and Security Rules (45 CFR Parts 160, 162, and 164);
- the Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology (45 CFR 170), hereafter referred to as the EHR Certification Rule;
- the Gramm-Leach-Bliley Act, Disclosure of Nonpublic Personal Information (Pub.L. 106-102, Title V, Subtitle A; codified at 15 USC, Subchapter I, Sec. 6801-6809), hereafter referred to as the Gramm-Leach-Bliley (GLB) Act; and
- the Privacy of Consumer Financial Information (16 CFR Part 313), referred to hereafter as the GLBA Financial Privacy Rule.

We chose to use only 10 legal statements to ensure that our study was short; a longer study may have discouraged some participants from taking the survey. When selecting the 10 legal statements to use for our study, we preferred legal statements that were shorter in length and ensured that each classification was exhibited by at least one cross-reference.

[2] www.ftc.gov/os/statutes/031224fcra.pdf

The only classification that does not appear in our study is the *incorrect* classification. We did not include an incorrect cross-reference because they require in-depth analysis to identify; we identify incorrect cross-references by observing footnotes in the legal text or analyzing the referenced legal text—neither of which we provided in our study. We employed the Qualtrics[3] survey software to deliver our survey.

B. Participants

The participants in our study are software engineers, healthcare professionals, legal domain experts, requirements engineers, software architects, development managers, individuals working in support and professional services roles, and other individuals that work in the field of healthcare IT. We targeted two organizations to recruit participants. First, we recruited participants from a trade association of 41 EHR vendors, including 9 of the top 10 EHR vendors based on number of providers attesting for MU Stage 1 incentives[4]. Second, we targeted a nonprofit consortium of 220 organizations that includes healthcare systems, healthcare IT vendors, academic medical centers, government agencies, and other technology vendors. Participants were recruited through email to various work groups and primary contacts at member organizations, and offered the chance to enter a random drawing to win a small gift card to a popular online book seller.

The expert group consisted of the author, a requirements engineering professor who is an expert in compliance requirements, a law professor who was a senior manager in the drafting of the HIPAA Privacy and Security Rules and the GLBA Financial Privacy Rule, and two senior PhD students who are familiar with compliance requirements research.

C. Research Methodology

We performed a pilot before running our full study. Current and former members of the Realsearch[5] and ThePrivacyPlace[6] research groups were recruited to participate in the study. The pilot ran for one week; after the pilot closed, we performed initial analysis on the results and collected verbal feedback from several pilot participants. Based on this analysis, we made no substantive changes to the survey instrument before running the full study.

We ran the full study by targeting the participant groups discussed in Section IV-B. The survey was open for 30 days; after the survey completed, we gathered the results from the Qualtrics software for comparison to the classification performed by the experts (mentioned in Section IV-B). Consensus was achieved on the expert results through discussion during a reading group meeting lasting two hours, involving the author, a requirements engineering professor, and two PhD students. When the expert group could not reach consensus, we set the cross-reference aside and consulted the law professor expert during a later call. The final expert consensus serves as the oracle against which we compare participant responses.

V. Participant Performance

In this section, we discuss participant demographics and performance and make observations based on our user study results.

A. Participant Demographics

We had 11 participants begin the pilot study and seven complete the pilot, that is, they completed every question. For the user study, we had 56 participants begin the survey and 33 complete it. We summarize the demographics of our user study in Table II. Note that the numbers may not add up to the number of participants, because participants may select multiple answers. For example, a participant may be both a healthcare practitioner as well as a business analyst.

Our survey included a free-form text box where participants could enter other roles they are currently in or held previously. In the pilot study, participants entered "software engineering researcher" and "software educator" for other roles. In the full study, participants entered "development manager", "program manager" (twice), "vice president product strategy", "infrastructure architect", "director of product management", and "quality manager" for other roles.

B. Participant Performance

Recall that participants were asked to classify ten legal statements using our cross-reference taxonomy. Cross-references can be classified with multiple classifications. For the ten legal statements used in our study, the experts classified each cross-reference with either one or two classifications. We say a selection *matches* between a participant and the expert group when the participant selects the same classification as the expert group. For example, if the experts classified a cross-reference as a constraint, a participant's selection would match if he or she also classified the cross-reference as a constraint. We say a participant has a *correct* answer when he or she makes the same number of selections as the experts and all of their selections match the expert selections. A participant response is *partially correct* if: (1) he or she selects fewer or more classifications than the expert group and one or more of their selections match the experts, or (2) he or she selects the same number of classifications as the expert group, and one selection matches and others do not. A participant response is *incorrect* if he or she selects no classifications that match the expert classifications. Table III displays the possible outcomes based on the number of classifications that are made by a participant and the expert group. We give correct answers one point, partially correct answers half a point, and incorrect answers zero points.

As displayed in Table III, there are different types of partially correct answers. We do not differentiate between these different types of partially correct answers in our scoring. Instead, we give each partially correct answer half a point. We selected this scoring rubric because there is no partially correct

[3] https://www.qualtrics.com/

[4] https://www.thehitcommunity.org/2012/05/hitc-data-watch-top-10-ehr-vendors-in-the-cms-ehr-incentive-program/

[5] http://www.realsearchgroup.org/realsearch/

[6] http://theprivacyplace.org/

TABLE II: Participant Demographics

		Pilot Study	User Study
		(# in role / median years of experience)	
Current Role	Business Analyst / Requirements Engineer	1 / 8	4 / 12.5
	Software Architect / Developer	3 / 3.75	17 / 15
	Quality Engineer / Tester	1 / 0	5 / 9.5
	Implementations / Support / Services	0 / 0	1 / 7
	Network Engineer / IT	0 / 0	1 / 3
	Compliance / Legal	1 / 0	3 / 3
	Healthcare Practitioner	0 / 0	2 / 7.5
	Other - S/W Eng. Researcher	1 / 3	0 / 0
	Other - S/W Educator	1 / 7	0 / 0
	Other - Dev. Manager	0 / 0	1 / 15
	Other - Program Manager	0 / 0	2 / 3
	Other - VP Product Strategy	0 / 0	1 / 12
	Other - Infrastructure Architect	0 / 0	1 / 3
	Other - Dir. of Product Management	0 / 0	1 / 2.5
	Other - Quality Manager	0 / 0	1 / 10
		(# in role / median years of experience)	
Previous Experience	Business Analyst / Requirements Engineer	2 / 5	5 / 4
	Software Architect / Developer	4 / 4.25	20 / 16
	Quality Engineer / Tester	0 / 0	6 / 8.5
	Implementations / Support / Services	0 / 0	8 / 3.5
	Network Engineer / IT	0 / 0	6 / 6.5
	Compliance / Legal	0 / 0	1 / 11
	Healthcare Practitioner	0 / 0	5 / 6
	Other - Compliance Consultant	0 / 0	1 / 11
	Other - Intelligence/Security	0 / 0	1 / 9
	Other - Healthcare Admin.	0 / 0	1 / 8
	Other - Dev. Manager	0 / 0	1 / 5
	Other - Clinical Informatics	0 / 0	1 / 12
	Other - Conf. Manager	0 / 0	1 / 10
	Other - S/W Support Services Manager	0 / 0	1 / 15
	Other - Policy Analyst	0 / 0	1 / 10
Median Years Experience Working a Highly Regulated Domain		1.75	12.5
Highest Education Level	High School Diploma / GED	0	3
	Associates	0	2
	Bachelors	1	21
	Masters	1	5
	PhD	5	0
	Professional Degree (M.D., J.D., etc.)	0	2

TABLE III: Possible User Study Outcome for a Single Question

		Experts Selected	
		One Classification	**Two Classifications**
Participant Selected	**One Classification**	1 match = correct	1 match = partially correct
			0 matches = incorrect
	Two Classifications	1 match = partially correct	2 matches = correct
		0 matches = incorrect	1 match = partially correct
			0 matches = incorrect

answer that is more compliant than another partially correct answer. In fact, the impact of a partially correct or incorrect classification on compliance depends on the context of the cross-reference. For example, if a cross-reference is actually unrelated but an engineer misclassifies it as a constraint, this error may lead to a more conservative interpretation because it over-constrains the software. However, if an engineer misclassifies the cross-reference as an exception, this error may lead to less compliant software.

Before we calculated participant performance, we set aside the participants who did not complete the entire exercise. We then totaled each participant's score based on our scoring rubric. The highest possible score for the exercise was ten points. We did not have enough participants for our data to be normal, so we use non-parametric statistical tests, specifically, we use Mann-Whitney U to test our hypothesis. The median pilot participant score was 7 out of a maximum of 10 points, while the median participant score for the full study was 5.5 out of a maximum of 10 points. Pilot participants spent a median time of 10 minutes, 53 seconds taking the survey, whereas participants taking the full study spent a median of 8 minutes, 7 seconds taking the survey.

As done in previous studies [12], we use a consensus approach to identify trends in participant responses; based on the degree of consensus among participant responses, we can measure which questions participants performed well on and which questions they struggled with. Table IV displays the participant responses by question, listed as a percentage of the total responses to the question. For each column in Table IV, we bold the classification that was made by the expert group. The percentages may not total 100, due to rounding. If participants achieved perfectly correct consensus on a question with a single classification, the expected outcome would be that 100 percent of the participant classifications match the expert classification for that question. For questions with two classifications, the expected outcome would be 50 percent of the participant classifications match the first expert classification and 50 percent match the second classification.

VI. Discussion and Observation

In this section, we discuss participant performance and make some observations about our study. Examining participant performance in Table IV, we see that participants performed well on three questions (Q3, Q4, and Q8), moderately well on two questions (Q2, Q7), and poorly on five questions (Q1, Q5, Q6, Q10).

Participants performed most poorly on question one, which states: "45 CFR 164.512(k)(3): A covered entity may disclose protected health information to authorized federal officials for the provision of protective services to the President or other persons authorized by 18 U.S.C. 3056 [...]" The expert group classified this as a constraint cross-reference but 39.4% of participants classified this as an unrelated cross-reference. We hypothesize that the participants and the experts had different assumptions about what requirements are unrelated to software. To test this hypothesis, we are planning on re-running our study with modifications. We will include a software specification in the participant materials; participants will be asked to classify a cross-reference as unrelated if it is unrelated to the software specification provided.

Participants appeared to perform better classifying definition cross-references than other types of cross references. They also appeared to perform poorly on questions with multiple classifications, only making one correct classification. As discussed in Section IV-A, we chose to use only 10 legal statements in our study to keep it short and encourage participants to take the survey. However, this small number of legal statements means that participants only saw one or two cross-references for each classification. Thus, we lacked the statistical power to test our observations above. In our future study, we plan on using a greater number of cross-references to gain statistical power to validate classification trends.

In the remainder of this section, we discuss our statistically validated observations.

1) Observation #1: Software Engineers are Not Well Equipped to Understand the Impact of Cross-References on Software Requirements: In Section IV, we introduced the null hypothesis for our user study:

> H_1*:* Individuals from the participant group have equivalent or greater precision than the expert classifications when classifying cross-references using the taxonomy.

In our study, we found that software practitioners had a median score of 5.5 for the exercise. Recall that the highest possible score for our exercise was 10 (the performance of the experts). Based on the Mann Whitney U test, we find that software practitioners cannot classify cross-references with the same precision as the expert group (p = 0.0002). Thus, we *reject* our null hypothesis, and conclude that software practitioners are *not* able to classify cross-references with equivalent or greater precision than the experts using our taxonomy. Cross-reference analysis is a subprocess of other compliance analysis [15]. If software practitioners are not well equipped to perform cross-reference analysis, they are likely not well equipped to perform other compliance analysis activities. Massey finds that software engineering graduate students are "ill-prepared to identify legally compliant software requirements with any confidence" [16]. We suggest that this is not unique to students but that software practitioners are ill-prepared as well.

Prior work has stated that software engineers need tools and techniques to help them address regulatory requirements [3], [8], [14], [16], [17]. Our work suggests that software engineers are likewise ill-equipped to use these tools and techniques even when they are provided to them. Previous studies have found that software engineers perform better on compliance tasks when provided tools and techniques (e.g., [14], [16], [17]) but to our knowledge we are the first computer science researchers to empirically study how well software practitioners perform compliance tasks when compared to experts.

In the next subsection, we compare the best and worst performers in the full study.

2) Observation #2: Participants with More Experience in Regulatory Domains Perform Better: Recall that, as part of our study, we collected various demographic characteristics for each participant, such as current role, past experience,

TABLE IV: Participant Responses by Question, as Percentage of Question Responses

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
Pilot Study										
Constraint	**71.4**	12.5	12.5	0.0	**28.6**	**11.1**	0.0	0.0	**75.0**	14.3
Exception	0.0	0.0	**87.5**	0.0	0.0	11.1	**50.0**	0.0	0.0	0.0
Definition	28.6	0.0	0.0	**87.5**	0.0	0.0	50.0	**87.5**	0.0	0.0
Unrelated	0.0	12.5	0.0	12.5	**42.9**	0.0	0.0	12.5	12.5	**42.9**
Incorrect	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
General	0.0	0.0	0.0	0.0	28.6	**66.7**	0.0	0.0	0.0	14.3
Prioritization	0.0	**75.0**	0.0	0.0	0.0	11.1	0.0	0.0	12.5	**28.6**
Full Study										
Constraint	**22.2**	5.9	7.9	3.0	**38.9**	**15.4**	2.2	5.4	**48.6**	17.1
Exception	19.4	8.8	**76.3**	0.0	0.0	10.3	**50.0**	0.0	8.1	0.0
Definition	11.1	11.8	2.6	**84.8**	16.7	7.7	30.4	**81.1**	16.2	5.7
Unrelated	36.1	5.9	0.0	6.1	**25.0**	5.1	8.7	2.7	13.5	**40.0**
Incorrect	0.0	2.9	2.6	0.0	2.8	0.0	2.2	0.0	2.7	0.0
General	5.6	5.9	7.9	3.0	11.1	**59.0**	2.2	10.8	8.1	22.9
Prioritization	5.6	**58.8**	2.6	3.0	5.6	2.6	4.3	0.0	2.7	**14.3**

education, and the participant's comfort level with reading and understanding software requirements and legal texts. To analyze the best performing and worst performing participants, we first sorted participants according to their performance. Once sorted, we examined the top-performing quartile and the bottom-performing quartile, comparing each of the demographic characteristics. We only determined one weak statistically significant trend among demographic characteristics that explain participant performance—participants with more experience in regulatory domains performed better than those with less regulatory experience (p = 0.0548).

3) Observation #3: Pilot Participants Performed Better than Software Practitioners: The pilot participants performed better than full study participants with (p = 0.0374) with median scores of 7 and 5.5 respectively. We hypothesize two possible reasons for this observation. First, the pilot participants were better educated than participants in the full study; in the pilot, 85.7% of the participants had an advanced degree (a masters, PhD, or professional degree), whereas in the full study, only 21% of the participants had an advanced degree. Second, the pilot participants came from ThePrivacyPlace and Realsearch research groups. These groups have a wide range of experience with compliance, security, and privacy requirements [1], [5], [7], [12], [17], [18]. This familiarity with the research field may have better equipped them to correctly classify cross-references.

VII. Threats to Validity

When designing any study, care should be taken to mitigate threats to validity. There are four types of validity that should be maintained: construct validity, internal validity, external validity, and reliability [19], [20]. Construct validity addresses the degree to which a case study is in accordance with the theoretical concepts used [20]. Internal validity measures the validity of cause-effect or causal relationships identified in a study [20]. External validity is the ability of a case study's findings to generalize to broader populations [20]. Finally, reliability is the ability to repeat a study and observe similar results [20].

Three ways to reinforce construct validity are: use multiple sources of reliable evidence; establish a chain of evidence; and have key informants review draft case study reports [20]. We employ multiple sources of evidence by selecting multiple regulations from two domains. To establish a chain of evidence, we maintained careful documentation when performing our analyses. Finally, our draft case study reports were reviewed by several members at ThePrivacyPlace.org as well as by a law professor who was a senior manager in drafting both the HIPAA Privacy Rule and the GLBA Financial Privacy Rule.

The studies described herein are exploratory in nature; we do not make causal inferences nor identify cause-effect relationships in our work. Because we make no causal inferences in our study, internal validity is not a concern [20].

Our results are currently applicable to the healthcare and financial regulations that we examined. Our future work will examine regulations and statutes in other domains to continue to reinforce our external validity. However, previous researchers outline anecdotal evidence that the HIPAA Privacy and Security Rule are similar to other regulatory texts [16]. Our study strengthens this anecdotal evidence, by demonstrating that techniques developed using healthcare regulations (not just the regulations promulgated pursuant to HIPAA) can be applied to other domains (namely, finance). To further strengthen external validity, we recruit participants from a variety of healthcare organizations through the trade groups and consortiums we targeted.

To reinforce our study's reliability, we carefully document our approach and make our study materials are available online[7].

VIII. Conclusion

In this paper, we present an empirical study that assesses the ability of software practitioners to understand the impact that legal cross-references have on their software. We found that software practitioners are not well equipped to understand the

[7] http://www4.ncsu.edu/ jcmaxwe3/CrossRefUserStudyMaterials.pdf

impact of legal cross-references. In addition, we discover that individuals with more experience in regulatory environments are better at classifying the impact of cross-references on software requirements than those with less experience. Our study suggests that requirements engineers should work with policy and legal domain experts when building software for highly regulated domains, as without these interactions they may lack the tools and insight to create compliant software. To the best of our knowledge, we are the first researchers to assess the ability of software practitioners to reason about compliance requirements in an industry setting.

This study lays the groundwork for future studies further examining how well software practitioners and healthcare IT professionals understand and apply regulatory requirements. Participants in our study spent a median of 8 minutes, 7 seconds taking the survey, meaning they spent no more than a few minutes reading the tutorial provided with the survey. We hypothesize that participants will do better with more training and feedback. We plan to conduct a follow-up study to test this hypothesis. In this future study, participants will again be provided a tutorial and a set of questions. After completing this set of questions, participants will be given feedback on their responses, then given a second set of questions. In addition, as discussed in Section VI, we will also provide participants with a software specification, to reduce ambiguity around what legal requirements may be unrelated to software systems. We will compare the participants' performance with the practitioners' performance in the study described herein.

Acknowledgments

This work is partially supported by NSF Grant #1217699. The authors thank the members of the Realsearch and ThePrivacyPlace research groups who participated in the pilot study as well as the software practitioners that participated in our final study.

References

[1] P. Otto and A. Antón, "Addressing Legal Requirements in Requirements Engineering," in *Proc. 15th Intl. Requirements Engineering Conf.* IEEE, 2007, pp. 5–14.

[2] Electronic Health Record Association, "Letter to the national coordinator for health information technology," 2012, http://www.himssehra.org/docs/20120731_EHRAssociationMUTimelineLetter.pdf.

[3] J. Maxwell and A. Antón, "Developing Production Rule Models to Aid in Acquiring Requirements from Legal Texts," in *Proc. 17th Intl. Requirements Engineering Conf.* IEEE, 2009, pp. 101–110.

[4] J. Maxwell, A. Antón, and P. Swire, "A Legal Cross-References Taxonomy for Identifying Conflicting Software Requirements," in *Proc. 19th Intl. Requirements Engineering Conf.* IEEE, 2011, pp. 197 – 206.

[5] J. Maxwell, A. Antón, P. Swire, M. Riaz, and C. McCraw, "A Legal Cross-References Taxonomy for Reasoning About Compliance Requirements," *Requirements Engineering Journal*, 2012.

[6] J. Maxwell, A. Antón, and P. Swire, "Managing Changing Compliance Requirements by Predicting Regulatory Evolution: An Adaptability Framework," in *Proc. 20th Intl. Requirements Engineering Conf.* IEEE, 2012.

[7] T. Breaux and A. Antón, "Analyzing Regulatory Rules for Privacy and Security Requirements," *IEEE Trans. on Software Engineering*, vol. 34, no. 1, pp. 5–20, 2008.

[8] S. Ghanavati, D. Amyot, and L. Peyton, "Compliance analysis based on a goal-oriented requirement language evaluation methodology," in *Proc. 17th Intl. Conf. on Requirements Engineering*. IEEE, 2009, pp. 133–142.

[9] D. Gordon and T. Breaux, "Reconciling Multi-Jurisdictional Requirements: An Case Study in Requirements Water Marking," in *Proc. 20th Requirements Engineering Conf.* IEEE, 2012.

[10] J. Cleland-Huang, A. Czauderna, M. Gibiec, and J. Emenecker, "A Machine Learning Approach for Tracing Regulatory Codes to Product Specific Requirements," in *Proc. 22nd Intl. Conf. on Software Engineering*. IEEE, 2010, pp. 155–164.

[11] T. Breaux, "Exercising Due Diligence in Legal Requirements Acquisition: A Tool-supported, Frame-Based Approach," in *Proc. 17th Intl. Requirements Engineering Conf.*, 2009, pp. 225 –230.

[12] A. Massey, B. Smith, P. Otto, and A. Antón, "Assessing the Accuracy of Legal Implementation Readiness Decisions," in *Proc. 19th Intl. Requirements Engineering Conf.* IEEE, 2011, pp. 207–216.

[13] J. Schmidt, A. Antón, and J. Earp, "Assessing Identification of Compliance Requirements from Privacy Policies," in *Proc. 5th Intl. Workshop on Requirements Engineering and Law*, 2012, pp. 52 –61.

[14] T. Breaux, "Legal Requirements Acquisition for the Specification of Legally Compliant Information Systems," Ph.D. dissertation, North Carolina State University, 2009.

[15] J. Maxwell, "Reasoning About Legal Text Evolution for Regulatory Compliance in Software Systems," Ph.D. dissertation, North Carolina State University, 2013.

[16] A. Massey, "Legal Requirements Metrics for Compliance Analysis," Ph.D. dissertation, North Carolina State University, 2012.

[17] J. D. Young Schmidt, "Specifying Requirements Using Commitment, Privilege, and Right (CPR) Analysis," Ph.D. dissertation, North Carolina State University, 2012.

[18] B. Smith, A. Andrew, M. Brown, J. King, J. Lankford, A. Meneely, and L. Williams, "Challenges for Protecting the Privacy of Health Information: Required Certification Can Leave Common Vulnerabilities Undetected," in *Proc. 2nd Annual Workshop on Security and Privacy in Medical and Home-Care Systems*. ACM, 2010, pp. 1–12.

[19] R. Feldt and A. Magazinius, "Validity Threats in Empirical Software Engineering Research-An Initial Survey," in *Proc. 22nd Intl. Conf. on Software Engineering and Knowledge Engineering*, 2010.

[20] R. Yin, *Case Study Research: Design and Methods*, 3rd ed., ser. Applied Social Research Methods Series. Sage, 2003.