MEASURING PATENT EXAMINATION CONSISTENCY AT THE UNITED STATES PATENT AND TRADEMARK OFFICE

GABRIEL COMENZO, HARMONI LARRABEE,

MARGARET MUNROE, REBECCA NORIS

December 9, 2021

TABLE OF CONTENTS

AUTHORSHIP

Section:	Author:
Abstract	Margaret Munroe
Executive Summary	Margaret Munroe
1. Introduction	All
2. Background	Gabriel Comenzo
2.1 Patent Application Process	Rebecca Noris
2.2 Patent Examination Process	Margaret Munroe
2.3 Types of Consistency	Margaret Munroe
2.4 Factors that Affect Consistency	Rebecca Noris and Margaret Munroe
2.5 Measuring Inconsistency in Decision-Making Fields	Harmoni Larrabee
2.6 Summary	Gabriel Comenzo
3. Methodology	Harmoni Larrabee
3.1 Industry Standards for Consistency Measurement Methods	Rebecca Noris
3.2 Effectiveness and Applicability of Consistency Measurement Methods	Margaret Munroe and Gabriel Comenzo
3.3 Compiling the Catalog	Harmoni Larrabee
4. Findings	Rebecca Noris
4.1.1 Factors that Cause Inconsistency at the USPTO	Rebecca Noris
4.1.2 Where Consistency Needs to be Measured at the USPTO	Gabriel Comenzo
4.1.3 Consistency Measurement Methods	All
4.2 Catalog of Consistency Measurement Methods	Margaret Munroe
5. Discussion	Harmoni Larrabee
5.1 Recommendations	Harmoni Larrabee
5.2 Limitations	Rebecca Noris
5.3 Future Research	Margaret Munroe
5.4 Our Experience	All
Appendix A: Literature Review of Consistency Measurement Across Industries	All
1. The FDA	Rebecca Noris
2. Court Judges	Harmoni Larrabee
3. Refugee Status Decisions	Harmoni Larrabee
4. Human Resources	Gabriel Comenzo
5. Insurance	Margaret Munroe
6. Healthcare	Margaret Munroe
7. Loan Lending	Gabe and Rebecca
8. Engineering	Margaret Munroe
Editor	Harmoni Larrabee

ACKNOWLEDGEMENTS

The project team would like to thank Martin Rater, and Dan Sullivan, our project sponsors at the USPTO. We would also like to thank Veronica Augburn-Seaforth, David Fitzpatrick, Elisabeth Foley, Dan Hunter, Andrew Koenig, Jeff Porter, Cathy Sias, and Robyn Sirkis at the USPTO for all of their help. Furthermore, we would like to thank all of the employees at the patent office who participated in our interviews. Finally, we would like to thank Professor Adrienne Hall-Phillips and Professor James Hanlan, our advisors on this project, for their support.

ABSTRACT

The purpose of this project was to provide the United States Patent and Trademark Office with methods for measuring consistency in the patent examination process. To fulfill this purpose, the team conducted a literature review of consistency measurement methods that are used across a variety of industries, then conducted interviews with employees at the patent office to gather their opinions on consistency. Based on the information collected in the literature review and interviews, the team developed a catalog in which those methods were tailored to meet the needs of the patent office. The catalog provided includes multiple consistency measurement methods tailored to the USPTO and information that will allow the patent office to choose the most appropriate method for any situation.

EXECUTIVE SUMMARY

The purpose of this project was to provide the United States Patent and Trademark Office with methods for measuring consistency in the patent examination process. To fulfill this purpose, the team first conducted a literature review of consistency measurement methods that are used across a variety of industries. The literature review included consistency measurement methods used in both governmental and nongovernmental organizations to measure consistency. These organizations included immigration agencies, the Food and Drug Administration, and the insurance and engineering fields, among others. Through research for the literature review, the team found that process maps, decision-making surveys, questionnaires, and benefit-risk frameworks are useful methods for measuring consistency. Several methods for quantifying the results of the above methods were also found, including percent agreement, six sigma, and various correlation coefficients.

Following the literature review, the team conducted interviews with employees at the patent office to gather their opinions on consistency. In interviews, process maps and decision-making surveys were thought to be effective by significantly more participants than other methods. However, the significant variation in participants' responses showed that all four of the methods could be useful for consistency measurement in the patent office.

Based on the information collected in the literature review and interviews, the team developed a catalog in which those methods were tailored to meet the needs of the patent office based on information provided by interviewees. A rubric that explains when to use each method, its effectiveness, and its advantages and disadvantages, is also included.

1. INTRODUCTION

The United States Patent and Trademark Office (USPTO) is a government agency in the United States Department of Commerce where inventors, companies, and institutions apply for patents and trademarks for their inventions and intellectual property. The USPTO's goal is to "Foster innovation, competitiveness, and economic growth, domestically and abroad, by providing a high-quality and timely examination of patent and trademark applications, guiding domestic and international intellectual property (IP) policy, and delivering IP information and education worldwide" (USPTO, 2021a).

To meet this goal, the USPTO must decide whether an applicant's invention or intellectual property fits the criteria that warrant a patent. The four main requirements are that an idea must be "statutory, novel, useful, and non-obvious" to qualify for a patent (Patentability Requirements Under U.S. Patent Law, 2019). Within the general requirements, there is a set of detailed standards the USPTO must confirm each invention meets before granting it a patent. Through a thorough examination, a patent is either approved or denied.

To provide a high-quality patent examination, the USPTO must ensure that these detailed standards are applied consistently to each patent application they receive. In 2020, over 8,000 patent examiners worked at the USPTO, and 653,311 United States patent applications were submitted (USPTO, 2020b). The high number of workers and high volume of applications create challenges in ensuring that all patent examiners evaluate patent applications consistently.

To improve the consistency of patent examination across patent examiners, the USPTO must explore the level and types of inconsistencies that exist. To better understand these issues, the USPTO requires methods to measure inconsistency among patent examiners.

This project developed a catalog of methods for measuring consistency among patent examiners at the USPTO. The catalog provides multiple methods for measuring consistency and a rubric for each method describing its effectiveness across various criteria. The USPTO can now choose methods from the catalog to integrate into their quality management system to measure consistency between patent examiners.

2. BACKGROUND

The team was tasked with recommending approaches to measuring inconsistencies within the patent examining process at the USPTO. To begin our analysis, we examined the patent application and examination process. Following this, we researched types of decision consistency and factors that affect consistency in decision-making processes. Subsequently, we felt it necessary to examine different methods to measure said inconsistencies across industries where decision-making is critical.

2.1 Patent Application Process

There are six steps that applicants must take before they can receive a patent. This is a different process from that of the patent examiner, although the two are connected. See Figure 1 for an overview of this process.

At the end of the process, if the invention does not meet the patentability criteria, the inventor will receive a notice of the ruling with an explanation. The inventor may legally appeal the decision or reapply after a patent is rejected (USPTO, 2021b).

If the invention is deemed patentable, the inventor receives a Notice of Allowance, which includes a record of any revisions made and a list of fees. Maintenance fees must be paid to maintain a patent, and if they are not, the patent expires (USPTO, 2021b).

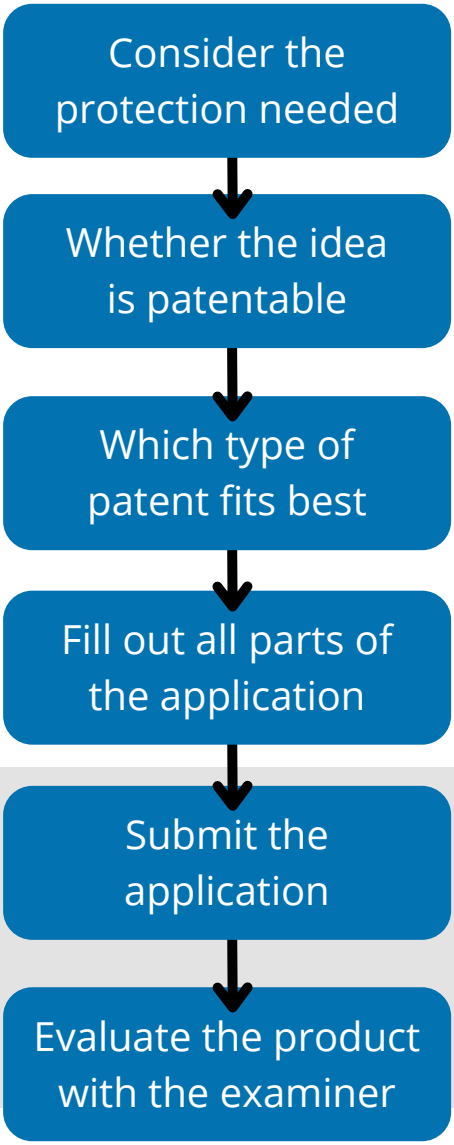

Figure 1: Steps in the Patent Application Process

2.2 Patent Examination Process

Throughout the patent examination process, the examiner reviews all of the sections of the application to ensure that each section meets all statutory requirements. The next stage is to search for the prior art related to the invention. In doing so, examiners must compare the new invention to previously patented inventions, thus ensuring the invention is sufficiently different from those previously patented. Figure 2 outlines all the steps of the examination process.

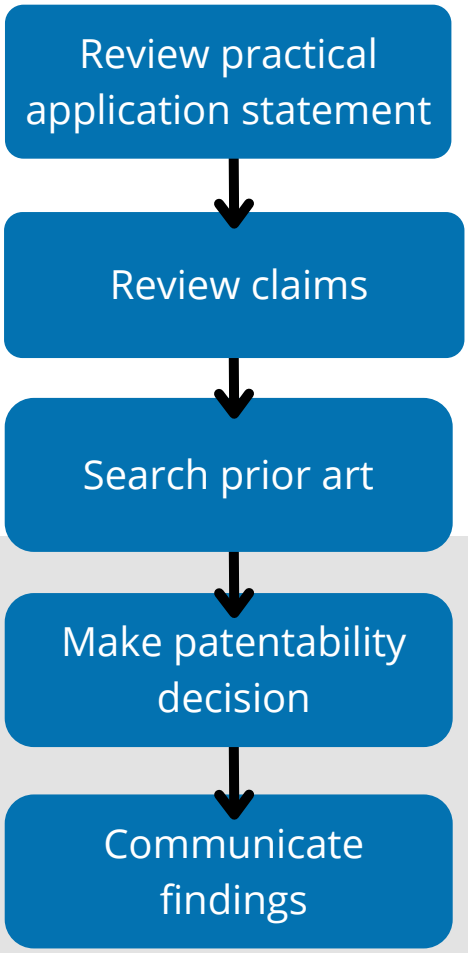

Figure 2: Steps in the Patent Examination Process

2.3 Types of Consistency

Three Types of Consistency:

- **Outcome Consistency**
- **Knowledge Consistency**
- **Process Consistency**

There are three related types of consistency that are important for both the USPTO and the other organizations that were researched. The first is outcome consistency, which is concerned with the decision made at the end of the process. Outcome consistency only assesses the outcome of a decision, not the process which was used to reach the decision (Rossmann, 2020).

The next type of consisntency is knowledge consistency. If two people do not share the same knowledge regarding their task, they cannot remain consistent with their task execution.

Finally, process consistency is concerned with whether the same steps are being taken to complete a process. This type of consistency requires real-time observation of the process, which can be collected either by an observing researcher or by the individual completing the process (Edwards et al., 2007).

2.4 Factors that Affect Consistency

Employee Engagement

To measure the consistency between employees, understanding what factors contribute to inconsistencies between their decision-making processes is vital. The first factor, employee engagement, plays a significant role in allowing a company to function effectively. One of the many definitions of employee engagement is "the emotional commitment the employee has to the organization and its goals" (Kruse, 2012). Studies have shown that employees' workloads have a significant impact on their levels of engagement. An increase in workload is tied to decreased employee engagement, which increases the likelihood of inconsistency, as less effort is expended on each task.

Experience Level

The second factor is the workers' level of experience. Research has shown that patent examiners' level of experience causes inconsistency in patent examinations. "We show that more experienced examiners cite less prior art, are more likely to grant patents and are more likely to grant patents without any rejections" (Lemley and Sampat, 2012).

Two groups within the patent office contribute most significantly to inconsistencies: those granting patents to undeserving inventions and those withholding patents from worthy inventions. Research has shown that the first group consists of senior examiners, who have at least five years of experience. The second group consists of junior examiners with less than five years' experience in the patent office (Tu, 2011, p. 2). The junior examiners' low approval rates result from their probation period; since their jobs are not stable, they do not want to take any risks by granting too many patents. Primary examiners, who have held their positions longer, have increased job security and are less strict about which patents they approve.

Finality

Another notable factor that impacts consistency is the lack of finality in patent rejection. The option for applicants to repeatedly submit the same patent application if the patent is not initially granted may eventually lead to their patent being accepted later. The cycle of resubmission undermines examination quality over time (Frakes and Wasserman, 2015). Since there is no limit to how many times inventors can resubmit their patents, this also exacerbates the application backlog.

As of 2012, about 40% of the backlog of patents awaiting examination were resubmitted applications (Frakes and Wasserman, 2015). To reduce the backlog, granting more patents is currently the examiners' best option.

2.5 Measuring Inconsistency in Decision-Making Fields

To better understand the problem of measuring decision-making consistency, the team researched methods of measuring consistency that are used across a wide range of industries. These industries included governmental organizations like the FDA, court judges, and refugee status decision-makers, as well as non-governmental organizations such as insurance, healthcare, human resources, loan lending, and engineering.

For the full literature review of consistency measurement across industries, see Appendix A.

2.6 Summary

The USPTO evaluates a large volume of patents, and there is a need for those evaluations to be executed consistently regardless of which examiner performs them. The nature of existing inconsistencies must be understood to improve consistency between patent examiners.

There are several types of consistency, but measurements of process and knowledge consistency are most effective in understanding the root of inconsistencies within decision-making processes. Many factors contribute to inconsistency in both general decision-making processes and the specific process of patent evaluation. In decision-based fields, it is a frequent practice to measure inconsistencies among employees' decision-making processes. As a result, there is a large body of existing literature on measuring consistency.

3.METHODOLOGY

The group developed a catalog of examples of consistency measurement methods across industries. The catalog was tailored for use by the USPTO to improve the consistency of the decision-making processes of the patent examiners. .

Three areas were considered in the development of a catalog of consistency measurement methods. The first was how other organizations and companies measure consistency. The second was which industry-standard methods of measuring consistency are most effective. The third was which methods are most applicable to the specific consistency concerns at the USPTO.

To gather the necessary information for this consideration, the team used research and interviews. The research enabled the team to determine the methods for measuring process, knowledge, and outcome consistency in various industries and which methods are most effective. Interviews with experts at the USPTO allowed the team to gather information about the inconsistency issues that the USPTO wants to address as well as what types of consistency measurement methods are feasible for the USPTO to use.

Research, interviews, and analysis of the information obtained through each were used iteratively to gather information about consistency measurement methods, their effectiveness, and their applicability to the USPTO. Then, the consistency measurement methods determined to be most effective and applicable were tailored for use at the USPTO and compiled into a catalog. See Figure 3 below for a diagram of the steps the team took through these methodologies to our deliverable.

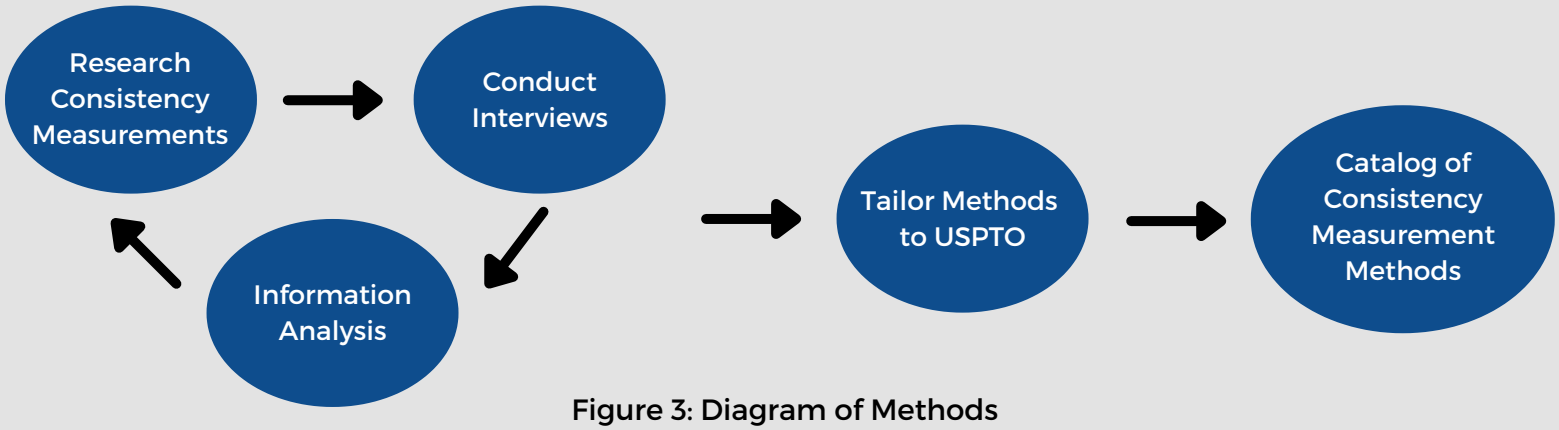

Figure 3: Diagram of Methods

The catalog includes tailored consistency measurement methods. The methods in the catalog are those that the team has determined to be most effective based on research of consistency measurement across industries and interviews with USPTO personnel.

For the full catalog of consistency measurement methods, see Appendix E.

The team is not responsible for implementing any of the methods outlined in the catalog. Instead, the USPTO's quality assurance team will implement the methods they see best fit.

3.1 Industry Standards for Consistency Measurement

To find examples of consistency measurement methods, the team looked at various organizations to find those with consistency measurement challenges similar to the USPTO. Each of the organizations in Figure 4 employs consistency measurement methods. The team conducted a literature review where in-depth information about the methods used at each organization is recorded.

Government Organizations	Non-Government Organizations
FDA Approval Decisions	Human Resources
Court Judges	Insurance Providers
	Healthcare
Refugee Status Decisions	Loan Lenders
	Engineering

Figure 4: Chart of Process-Oriented Fields

3.2 Effectiveness and Applicability of Consistency Measurement Methods

Once the team had compiled information about standard methods of comparing consistency in selected fields, interviews were conducted to determine which methods are most effective and applicable at the USPTO.

First, interviews were conducted with senior leaders. The purpose of these interviews was to get a broad sense of the consistency concerns at the USPTO as well as to get senior leaders' opinions on specific consistency measurement methods. Interview questions included factors that the senior leaders believe affect consistency of patent examiners' decisions and their thoughts on the effectiveness of specific consistency measurement methods (see Appendix B).

Interviews with patent examiners' supervisors were used to determine which consistency measurement methods align most with the supervisors' concerns. Participants were asked about factors that they believe contribute to inconsistencies in the patent examining process and places in the patent examining process where consistency measurement can be applied. Participants were also asked for their opinions on the effectiveness of specific consistency measurement methods (see Appendix C).

Finally, interviews with quality assurance personnel were used to determine which measurement methods are feasible to implement at the USPTO. The questions in these interviews focused on which consistency measurement methods would work well within the USPTO's existing quality assurance framework (see Appendix D).

All interview responses were coded using a standard semi-structured interview coding procedure. The coded responses were used to determine which of the consistency measurement methods found in the literature review best suit the USPTO's needs.

As the team learned more about which methods are most applicable at the USPTO by analyzing interview responses, these methods were researched further. As the interviews produce information regarding which decision-making factors cause the most inconsistency at the USPTO, subsequent research into each method's effectiveness in measuring each factor was executed. For a chart of these decision-making factors, see Figure 5 below. Through iterative interviews and research, the team determined the effectiveness and applicability of each consistency measurement method at the USPTO.

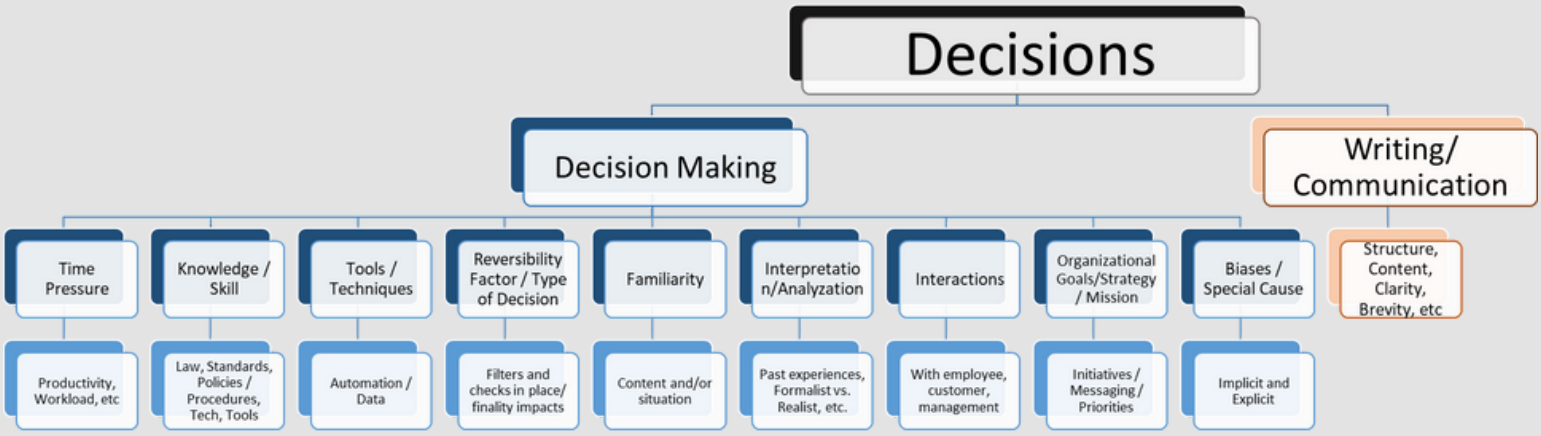

Figure 5: Chart of Factors Affecting Decision-Making

3.3 Compiling the Catalog

The team collected the necessary information about consistency measurement methods, and compiled a catalog of nine of the most effective and applicable methods.

For each method in the catalog, four pieces of information have been enumerated. First, a general description of the method is given. Then, a version of the method that has been tailored for use at the USPTO is provided. Next, a rubric that ranks the method's effectiveness on several criteria is provided (see Figure 6). Finally, references to examples of the method being applied in other industries are given.

To tailor each method for use at the USPTO, the team has provided specific suggestions on how to implement the method with content specific to the USPTO. The suggestions were developed based on the team's research of the respective method and information from interviews about the method.

Factors	Type of Consistency	Applicability
The decision-making factors that the method measures	The type of consistency that the method measures	The method's applicability at the USPTO
Accuracy	**Maturity**	**Other**
The accuracy of results obtained by the method	Whether the method is best used for well-established decision-making processes or developing ones	Advantages & disadvantages of the method which are not covered by other categories

Figure 6: Rubric for Rating Consistency Measurement Methods

4. FINDINGS

4.1 Factors that Cause Inconsistency at USPTO

Many participants discussed working style, time management skills, personality, interpretation of policies, and attention to detail as the main examiner characteristics that impact consistency. These factors can impact how an examiner approaches a problem, and the differences in approach can lead to different outcomes. Participants also mentioned external factors, both personal or work-related, such as the time of year, past experiences while working, any external personal factors, and time of year. One participant explained further that at certain times of the year, examiners may have a larger workload than other times due to a time crunch to meet their quotas or other deadlines.

Participants also said that an examiner receiving a patent application outside their area of art or expertise is likely to increase inconsistency. Training was also mentioned as a factor, with participants saying that examiners may interpret their training differently than one another, and that additional legal training for examiners might be needed to decrease inconsistency.

Other responses regarding knowledge discrepancies included shifts in technology, changing laws, lack of understanding of systematic impacts, and the amount of background work an examiner does before reading claims in a patent application.

There are also factors that cause inconsistencies within the examination process. One factor that interviewees mentioned repeatedly was meeting quotas. One participant said, “With finals, you have a smaller amount of time to write them up and there may be instances where you need to spend more time on that and that’s going to make you less consistent in your production.” Other factors that participants named included the way that a case is brought into the office, if there are any inconsistencies or translation issues in the application, and the necessary actions taken for different areas of art. There is also a lot of flexibility and subjectivity within the examination process. Participants’ responses indicated that each case is different and, depending on how the examiner interprets it, they will search for information, use tools, and make decisions as they see fit.

Participants further discussed how they type of patent being evaluated may impact consistency. Many of the responses were related to the patent's area of art. For example, one interviewee said that what goes into the examination of a chemistry patent application is going to be different than what goes into examining an engineering patent as each has different standards and laws. Additionally, poorly written patent applications, cultural differences, and patents that are difficult to categorize into the right area of art were named as causes of inconsistency.

Overall, there are many factors that cause inconsistency in the USPTO across all four categories that were asked about in interviews. However, interviewees mentioned more factors in some categories than others. Examiner characteristics had the most mentions, followed by the examination process, and lastly knowledge discrepancies and types of patents (see Figure 7). The categories with more mentions of factors may be most important to investigate to improve consistency in the patent examination process.

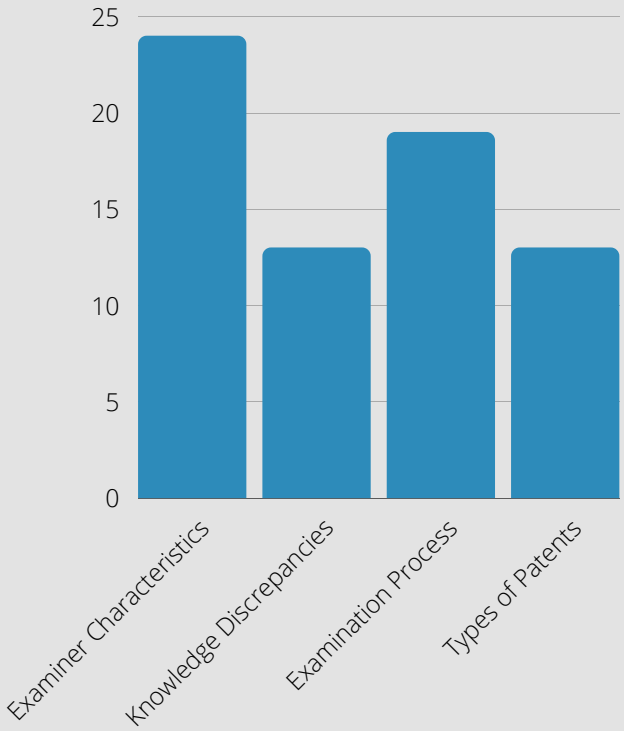

Figure 7: Number of Mentions of Factors That Cause Inconsistency

4.2 Measuring Consistency

One common theme among interview responses was that outcome consistency is particularly important to measure. Many participants stated that it is important for patent examiners to be consistent when determining if an application is patentable or not. One interviewee said, “The consistency of the final decision itself [is important to measure]. When an application is submitted by an applicant, they are in good faith that the application will be reviewed fair and equal to any other patent. If they lose this faith, then we are discredited as an organization.”

Another common theme in responses was the importance of measuring consistency in the early phases of the patent examination process. One interviewee said that if "early steps in the patent examination process are done incorrectly, the effects can trickle down to the rest of the process and potentially influence the final decision."

Other interviewees mentioned a wide variety of places within the examination process where it is important to measure consistency (see Figure 8). A quarter of participants said that measuring consistency in every part of the process is important. Overall, it is clear that patent office employees feel that consistency needs to be measured throughout the patent examination process.

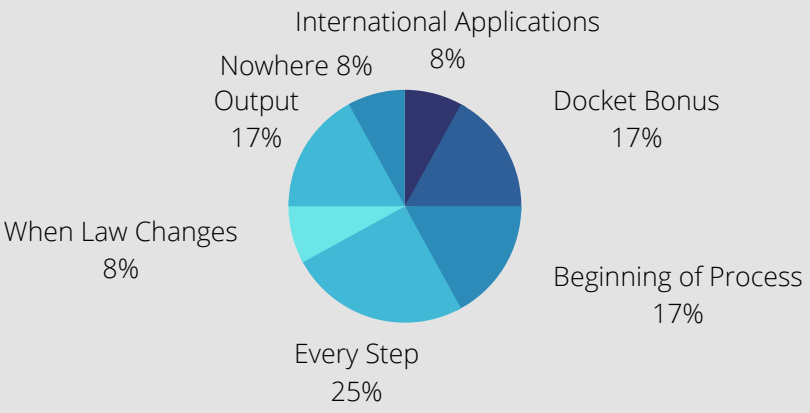

Figure 8: Where in the Patent Examination Process it is Most Important to Measure Consistency

4.3 Consistency Measurement Methods

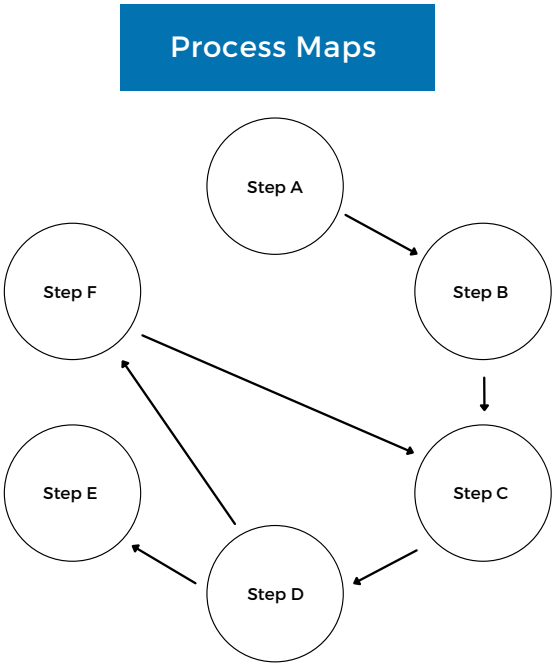

Figure 10: Process Map Example

The first method discussed during interviews was process mapping. A process map is a diagram showing all the steps needed for a particular process. The diagram is given to participants, who are asked to draw arrows to indicate the order in which they complete the steps of the process. Then, the diagrams can be compared to find inconsistencies in the ways that the different participants complete the process. See Figure 10 for an example of a process map.

Process maps received generally positive reviews from patent office personnel, with 61.54% of participants ranking it first or second compared to the other methods.

The majority opinion among interviewees was that these maps would be effective for consistency measurement within the examination process.

About a third of respondents believed that each examiner's map would look vastly different. If this is the case, it will be important for those who examine the resulting process maps to determine which of these variations are allowable and which should be eliminated. However, there were also several respondents who thought every process map would come out nearly the same. They asserted that there is variation that the maps would not be able to show.

For example, one participant stated that, even if two examiners complete the same steps in the same order, they could do the individual steps differently, and those differences would not be visible on a process map. Several interviewees also expressed concerns about the difficulty of development and implementation of the maps.

There was no consensus on where in the patent office process maps could be used, suggesting that process maps could be adapted and applied across nearly every part of the patent examination process. The areas where interviewees were most interested in using process maps included the examiner level, supervisory level, and quality assurance. Some participants specified that they would be interested in seeing maps of the early steps of the examination process, how examiners put their findings into writing, and high-level processes including application routing.

Participants raised questions about how to decide which steps and resources to show on the map template. These questions are best answered on a case-by-case basis, as there are several groups within the patent office that could use this method, and each group may have their own unique process map template.

One respondent summed up benefits and drawbacks of this method, saying, "I think there could be a lot of valuable information in there, if I'm not the one responsible for figuring out how to make a comprehensive map for somebody."

While there were both positive and negative opinions of process maps among those interviewed, process maps were ranked second-most effective overall by the interviewees out of the four methods discussed. They are better suited than the other methods for measuring process consistency, as many of the others focus on knowledge or outcome consistency.

Decision-Making Surveys

The next method presented in interviews was decision-making surveys. Decision-making surveys are a specific type of survey where a series of fictional scenarios are presented to survey participants. The participants are each asked how likely they would be to make a given decision in each fictional scenario. A Likert-type scale is provided for the participant to give their answer. Participants are then asked to provide reasoning for their answers. Responses are collected anonymously and compared to find inconsistencies between different participants' answers to the same scenarios.

46.15% of participants stated that they believe this method would yield useful information about consistency at the USPTO. As one participant said when asked for their opinion on decision-making surveys, it "would be fascinating […] to see how inconsistent we would be as an office in making a decision." About one-third of participants also mentioned that there is precedent of using this method or a similar method to assess examiner consistency at the patent office.

Several participants also noted that, if the fictional scenarios were constructed accordingly, this method could be used for decisions throughout the patent examination process, not just for the end decision of approval or rejection.

However, some participants also expressed concerns about the effectiveness of decision-making surveys. Two participants stated concerns that patent examiners may give biased answers on this type of survey. Another two participants mentioned that this type of survey would only be useful if the types of patents in the scenarios were specified to match the area of art in which the examiner taking the survey specializes.

Of the four consistency measurement methods included in the interviews, decision-making surveys received the most positive feedback from interview participants overall. Overall, decision-making surveys were ranked most effective by interviewees.

Benefit-Risk Framework

The third consistency measurement method that was described during the interviews is a benefit-risk framework. A benefit-risk framework is a method to weigh the benefits and the risks associated with a decision. To use this framework, participants record all the benefits and risks they consider while making a particular decision.

Participants' opinions about this method were split, with 6 participants saying that it would be useful and 6 participants saying that it would not be relevant to the examination process.

One participant shared, “I have used them before […] in other agencies and it’s a very easy tool to use, very easy to deploy, and people understand it very quickly.” In contrast, a different interviewee said, "I’m not thinking of too many examples of where I could see that being as relevant for the examination process.”

The participants who liked the benefit-risk method responded that they could see it being used in several places in the patent examination process, including the first office action, in the patentability determination stage, or throughout the entire process. A few responses also indicated benefit-risk frameworks could be used at the management level.

Many interviewees expressed that they could not picture how benefits and risks are relevant to the examination process. However, just as many participants thought that if benefit-risk frameworks were implemented correctly, they could help to identify where there is inconsistency in the patent examination process.

Questionnaires

The final consistency measurement method addressed in interviews was questionnaires. A questionnaire is a series of questions, usually with multiple choice options or a Likert scale to provide answers. Some questions may also be open-ended. Participants' responses can be compared to find inconsistencies in participants’ answers.

Responses to this method varied greatly. Some participants thought that questionnaires are effective, while others did not. Several interviewees expressed concerns that questionnaires are not effective unless they are constructed very precisely. Some participants also noted that questionnaire responses are often difficult to categorize. A few interviewees said that using multiple choice rather than open-ended questions may make this method more effective.

When interviewees were asked where it would be best to implement questionnaires, they gave a wide range of responses. Responses included the beginning of the patent examination process, testing new examiners, examiners’ work style, legal knowledge, and the end of the patent examination process.

The wide range of responses suggest that questionnaires could be applied to consistency measurement in almost any step of patent examination. Of the four consistency measurement methods presented in interviews, questionnaires were ranked lowest in effectiveness. However, those that did think questionnaires could be effective for consistency measurement thought that they had a wide range of applications.

Qualitative vs. Quantitative

One interview question asked whether qualitative or quantitative methods are better for measuring consistency. Many participants said that, while quantitative methods are important for identification of problems and measurement of their true impacts, qualitative methods are useful for getting to constructs that cannot be quantified. The responses to this question (see Figure 11) show a preference for quantitative methods, but they also show that the employees of the patent office feel that a combination of quantitative and qualitative methods is most effective.

One interviewee explained it by saying, "I'm not sure how we would separate them . . . focusing on one sometimes detracts from the value of the other, so we need both."

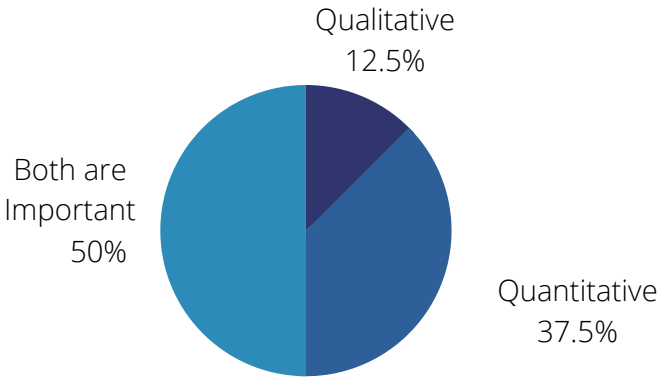

Figure 11: Qualitative vs. Quantitative Methods

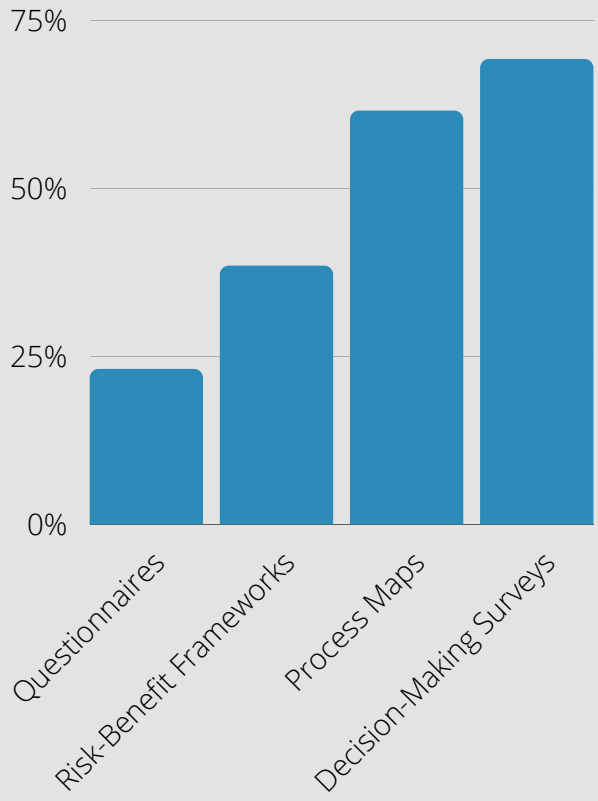

Figure 12: Interviewee Preference for Each Method

Additional Findings

Each of the four consistency measurement methods presented in the interviews received mixed feedback. However, there was a clear overall preference for process maps and decision-making surveys for measuring consistency in patent examination. See Figure 12 for a breakdown of the percent of interviewees who ranked each method either most effective or second-most effective of the four methods.

5. DISCUSSION

Every employee at the USPTO has different opinions about which kinds of consistency need to be measured and where consistency needs to be measured in the patent examination process. Employees also have widely varying opinions about which methods of measuring consistency would be effective. Overall, process maps and decision-making surveys were rated most effective by the USPTO employees we interviewed, however individual opinions about those methods varied greatly. The lack of consensus among interviewees suggests that there is no one method that is most effective for measuring consistency at the USPTO. So, this project produced a catalog of many different methods for measuring consistency, which includes the advantages and disadvantages of each method and specific recommendations on where and how to use each method.

5.1 Recommendations

The specific consistency measurement methods that the team recommends for measuring consistency in patent examination can be found in the catalog in Appendix E. In order to most effectively choose a method from the catalog, the team recommends using the following procedure.

In interviews, participants gave widely varied responses about the kinds of consistency that need to be measured at the USPTO. The responses covered the whole range of factors related to consistency and steps in the patent examination process, suggesting that nearly every type of consistency at every step in the patent examination process may be worthwhile to measure. So, to use the catalog effectively, the team recommends first choosing the type of consistency that is to be measured, then choosing where in the patent examination process consistency measurement will occur and which consistency-related factors the measurement will assess.

Once the above factors have been determined, the catalog of consistency measurement methods can be consulted to determine which methods are most applicable for the specific consistency measurement that is desired. The rubrics corresponding to each method in the catalog allows for easy determination of whether each method is applicable to a particular situation. The information in each rubric includes the method's effectiveness for measuring specific types of consistency, specific consistency-related factors, consistency in specific parts of the patent examination process, and more. The rubrics are designed to aid selection of the most applicable method for measuring any given form of consistency.

5.2 Limitations

One of the limitations of this project is that the results the team found are limited to the industries researched. The consistency measurement methods included in the catalog are not an exhaustive list and cannot be generalized across all industries. Also, the duration of the project was only seven weeks, which isn't enough time to look at every industry and method for measuring consistency.

Furthermore, because patent examiners are unionized, this project was unable to interview them, although their input would have been useful.

Lastly, because of the Covid-19 pandemic, the patent office is working remotely, so the project team could not work at the patent office.

5.3 Future Research

As a result of the limitations discussed in the previous section, there is more research that could be done in the future. Primary and Secondary patent examiners could be interviewed about their consistency experiences. This would provide an even closer view of the examination process than this project's work with Supervisory Patent Examiners.

While the catalog created in this project will be very useful to the USPTO, the project team was not able to be involved in the implementation of any of the methods in the catalog. Implementation could potentially be done by a future IQP team, especially since Martin Rater and others at the patent office are planning to work on consistency much more in the coming years.

5.4 Our Experience

Gabe

Over the past 7 weeks, I have had many valuable experiences and learned new skills that can translate to the workforce. First off, I found out what it is like to work full time. 9 am-5 pm five days a week took a little bit to get used to, but once you get into a routine, it gets easier.

Second off, I improved my social skills with people. Whether it was hosting meetings and interviews with staff at the USPTO or just working in a group, I feel as if my skills of interacting with people have improved. Overall, this experience is something I would not change one bit.

Maggie

I tried to spend just as much time seeing the city as working on the project, and it was a great city to be in, with plenty of history but also plenty of more modern stuff to do. It was cool to live in a city, I've been in Worcester for a couple years but DC had more places to go and more going on.

It was also good to get exposure to a regular working week, but it wasn't bad because I enjoyed the work. Working with the patent office also gave me some ideas of a future job, I wasn't really sure before because a math major is pretty broad, but I got some ideas from working with data scientists and others.

Rebecca

Being in D.C. and working on this project for the past 7 weeks has been a great experience that I have learned a lot from. Working 5 days a week, similar to a full time job, was a cool experience given most of the work I've done previously has been in a school setting. I really enjoyed interacting and working with others like staff the USPTO, my project team members, and the other students on this IQP. I was also able to work on my skills including researching, writing, and presenting.

I also really enjoyed the opportunity to travel to and explore D.C. especially since I have never been here before. It was interesting being able to go to museums and sight-seeing.

Harmoni

Over the course of this project, I've learned a lot about the way that the patent office operates and what it takes to get a patent. I also learned a lot about the many types of consistency, and I now have a new perspective on every process I encounter.

Living and working in Washington, DC for the past several weeks has also allowed me to learn more about the history of the United States and see historical monuments and museums.

BIBLIOGRAPHY

Ahmad, N., & Qahmash, A. (2020). Implementing Fuzzy AHP and FUCOM to evaluate critical success factors for sustained academic quality assurance and ABET accreditation. PLoS ONE, 15(9), e0239140. https://link.gale.com/apps/doc/A635747804/AONE?u=mlin_c_worpoly&sid=bookmark-AONE&xid=b60bf511or

Angelis, A., & Phillips, L. D. (2020, June 11). Advancing structured decision-making in drug regulation at the FDA and EMA. British Pharmacological Society | Journals. https://bpspubs.onlinelibrary.wiley.com/doi/full/10.1111/bcp.14425.

Bowman, E. H. (1963). Consistency and optimality in managerial decision making. Management Science, 9(2), 310–321. http://www.jstor.org/stable/2627409

Brass, E. P., Lofstedt, R., & Renn, O. (2011). Improving the decision-making process for nonprescription drugs: a framework for benefit–risk assessment. Clinical Pharmacology & Therapeutics, 90(6), 791-803.

Bryce, C., Webb, R., Cheevers, C., Ring, P., & Clark, G. (2016). Should the insurance industry be banking on risk escalation for solvency II? International Review of Financial Analysis, 46, 131–139. https://doi.org/10.1016/j.irfa.2016.04.014

Byun, J., Rhew, S., Hwang, M. et al. (2014). Metrics for measuring the consistencies of requirements with objectives and constraints. Requirements Eng, 19, 89–104. https://doi-org.ezpv7-web-p-u01.wpi.edu/10.1007/s00766-013-0180-9

Cabantous, L., Hilton, D., Kunreuther, H. et al. (2011). Is imprecise knowledge better than conflicting expertise? Evidence from insurers' decisions in the United States. Journal of Risk and Uncertainty, 42, 211–232. https://doi-org.ezpv7-web-p-u01.wpi.edu/10.1007/s11166-011-9117-1

Calandro, J., Jr. (2006). Accident year development, bonus banks, and insurance incentive compensation. Risk Management and Insurance Review, 9(2), 205. http://ezproxy.wpi.edu/login?url=https://www.proquest.com/scholarly-journals/accident-year-development-bonus-banks-insurance/docview/209610926/se-2?accountid=29120

Carey, M. (2002). Some evidence on the consistency of banks' internal credit ratings. Credit ratings: Methodologies, Rationale and Default Risk, London (Risk Books).

Cohen, A., Klement, A., & Neeman, Z. (2015). Judicial decision making: A dynamic reputation approach. The Journal of Legal Studies, 44(S1), S133–S159. https://doi.org/10.1086/682689

Collins, P. M. (2008). The consistency of judicial choice. The Journal of Politics, 70(3), 861–873. https://doi.org/10.1017/s002238160808081x

DeBell, L. E., & Chesney, D. L. (1982). The FDA inspections process. Food, Drug, Cosmetic Law Journal, 37(2), 244–249. http://www.jstor.org/stable/26658665

Duchessi, P., Shawky, H., & Seagle, J. P. (1988). A Knowledge-Engineered System for Commercial Loan Decisions. Financial Management, 17(3), 57–65. https://doi.org/10.2307/3666072

Edwards, K., Jensen, K. L., Haug, A. (2007). Measuring process and knowledge consistency: A necessary step before implementing configuration systems. Innovative Processes and Products for Mass Customization, 77-88. GITO.

Frakes, M. D., & Wasserman, M. F. (2015). Does the U.S. Patent and Trademark Office grant too many bad patents?: Evidence from a quasi-experiment. Stanford Law Review, 67(3), 613-676. http://ezproxy.wpi.edu/login?url=https://www-proquest-com.ezpv7-web-p-u01.wpi.edu/scholarly-journals/does-u-s-patent-trademark-office-grant-too-many/docview/1664233628/se-2?accountid=29120

Howard, R. M. (2005). Comparing the decision making of specialized courts and general courts: An exploration of tax decisions. The Justice System Journal, 26(2), 135–148. http://www.jstor.org/stable/27977228

Justia. (2019, June). Patentability requirements under U.S. patent law. https://www.justia.com/intellectual-property/patents/patentability-requirements/

Kruse, K. (2012, June 22). What is employee engagement. Forbes. https://www.forbes.com/sites/kevinkruse/2012/06/22/employee-engagement-what-and-why/?sh=7cedcaf77f37

Kürzinger, M.-L., Douarin, L., Uzun, I., El-Haddad, C., Hurst, W., Juhaeri, J., & Tcherny-Lessenot, S. (2020). Structured benefit–risk evaluation for medicinal products: review of quantitative benefit–risk assessment findings in the literature. Therapeutic Advances in Drug Safety. https://doi.org/10.1177/2042098620976951

Lemley, M., & Sampat, B. (2012). Examiner characteristics and patent office outcomes. The Review of Economics and Statistics, 94(3), 817-827. http://www.jstor.org/stable/23261480

McHugh M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Norris, S. (2019). Examiner inconsistency: Evidence from refugee appeals. (Becker Friedman Institute for Economics Working Paper No. 2018-75) http://dx.doi.org/10.2139/ssrn.3267611

Pärnamets, P., Tagesson, A., Wallin, A. (2020). Inconsistencies in repeated refugee status decisions. Journal of Behavioral Decision Making, 33(5), 569-578. https://doi.org/10.1002/bdm.2176

Reidy, T. J., Silver, R. M., & Carlson, A. (1989). Child custody decisions: A survey of judges. Family Law Quarterly, 23(1), 75–87. http://www.jstor.org/stable/25739798

Rejikumar, G., Aswathy Asokan, A., & Sreedharan, V. R. (2020). Impact of data-driven decision-making in Lean Six Sigma: an empirical analysis. Total Quality Management & Business Excellence, 31(3-4), 279-296.

Rosman, A. J., & Bedard, J. C. (1999). Lenders' decision strategies and loan structure decisions. Journal of Business Research, 46(1), 83-94.

Rossmann, F. (2020). Decision quality vs. decision outcome: How do you know you've made a good decision? Medium. https://medium.com/agileinsider/decision-quality-v-decision-outcome-the-often-neglected-role-of-luck-and-bad-luck-in-decision-d398065be677

Schober, Patrick, et al. (2018). Correlation coefficients: appropriate use and interpretation. Anesthesia & Analgesia, 126(5), 1763–1768. https://doi.org/10.1213/ane.0000000000002864

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research, and Evaluation, 9(1), 4

Tu, S. (2011). Luck/unluck of the draw: An empirical study of examiner allowance rates. Stanford Technology Law Review, 10. http://dx.doi.org/10.2139/ssrn.1939508

United States Food and Drug Administration (FDA). (2021, March 30). Inspection Guides. U.S. Food and Drug Administration. https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/inspection-references/inspection-guides

United States Patent and Trademark Office (USPTO). (2020a). 2103 patent examination process. United States Patent and Trademark Office – An Agency of the Department of Commerce. https://www.uspto.gov/web/offices/pac/mpep/s2103.html

United States Patent and Trademark Office (USPTO). (2020b). U.S. patent statistics chart calendar years 1963 - 2020. U.S. Patent Statistics Summary Table, Calendar Years 1963 to 2020, 05/2021 update. https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm

United States Patent and Trademark Office (USPTO). (2021a). About us. United States Patent and Trademark Office - An Agency of the Department of Commerce. https://www.uspto.gov/about-us

United States Patent and Trademark Office (USPTO). (2021b). Patent process overview. United States Patent and Trademark Office - An Agency of the Department of Commerce. https://www.uspto.gov/patents/basics/patent-process-overview#step6.

van Koppen, P. J., & Kate, J. T. (1984). Individual differences in judicial behavior: Personal characteristics and private law decision-making. Law & Society Review, 18(2), 225–247. https://doi.org/10.2307/3053403

Watt, I. (1997). Basic differential diagnosis of arthritis. Eur Radiol 4(7), 344–351. https://doi-org.ezpv7-web-p-u01.wpi.edu/10.1007/s003300050165

APPENDIX A

Literature Review of Consistency Measurement Across Industries

1. The FDA

One consistency measurement method used by the U.S. Food and Drug Administration (FDA) is a benefit-risk framework (BRF). The BRF that the FDA uses was recently developed to "improve the clarity and consistency in communicating the reasoning behind the FDA's decisions" (Angelis & Phillips, 2020). The BRF that the FDA currently uses is entirely qualitative. Angelis and Phillips' paper discusses the argument that FDA should implement quantitative modeling backed by decision theory to complement their current BRF and improve decision-making for US drug regulations.

The second consistency measurement method used by the FDA is a list of guides they use as a baseline for product regulations. There is a guide for each type of product the FDA approves: biotechnology, computer issues, devices, drugs, foods, cosmetics, and miscellaneous products. Each guide is a list of criteria that a product must meet to be approved by the FDA. These criteria help control the level of consistency in decisions on whether to approve a product or not. As an example, under the devices category, there is an inspection guide for medical device manufacturers. This guide includes substantial information about records, equipment, inspection standards, and personnel for the given subject. This is the reference material that FDA personnel use when evaluating a product for approval. However, the FDA's website also states that alternative approaches may be taken if applicable and these documents "do not bind FDA and do not confer any rights, privileges, benefits or immunities for or on any person(s)" (FDA).

A third method used by the FDA to ensure consistency in the development of a standard procedure to follow for all inspections. The inspection process involves two or more people, including a lead investigator and an individual with the training that is necessary for the task (DeBell & Chesney, 1982). In addition, all the investigators have at least a bachelor's degree, some have work experience in other fields, and they all have extensive experience which they receive on the job (DeBell & Chesney, 1982). The inspections are usually conducted as part of routine checks, special assignments such as a recall, or a follow-up to a complaint.

The procedure for inspections has four main steps. First, background information such as previous files and inspection reports related to the product being reviewed. Each of the firms being inspected must be made aware of the FDA's inspectional authority and the investigator (DeBell & Chesney, 1982). Next, the inspection usually consists of a general tour of the facility where the investigator takes pictures and collects samples of product labeling and copies of any relevant records to provide a thorough analysis of the firm. Then, the investigator provides

details of their observations with any corrections that need to be made by the firm. Finally, a narrative report on the investigation is submitted to the investigator's supervisor along with a list of observations (Form FD-483) and other documentation collected during the investigation (DeBell & Chesney, 1982).

2. Court Judges

There are two main methods of consistency measurement used to assess consistency in judges' decisions. The first method is a statistical analysis of existing data (precedent) about judges' decisions in past cases. The goal of this analysis is to find patterns that indicate the level of inconsistency in the judges' decisions or correlation between inconsistency in the judges' decisions and the factors of interest of the study. Statistical analysis methods tend to be used in studies where the primary interest is inconsistency in decision outcomes, not in the decision-making process. Three examples of the use of statistical analysis to measure consistency in judges' decision-making will be discussed here.

In one study, statistical analysis was used to investigate the differences in decision-making between general and specialized courts on tax-related cases (Howard, 2005). In this study, there was a particular interest in determining whether systematic differences in expertise and ideology between the two types of courts had an impact on the decisions made by those courts. The study analyzed data on tax-related cases that were handled by the specialized tax court and the general district courts in 1996 and 1997. The study created a set of variables related to the ideology of the judges, the structure of the court that handled the case, and the local versus national outlook of the court that handled the case. For each case, archived data was used to determine a value for each of the variables. Some of the variables included the outcome of the case (whether the judge ruled for or against the defendant), whether the judges had experience working for the IRS, whether the defendant hired an attorney, the political ideology of the judge, and the type of tax case.

The data for each of the variables across cases were aggregated for the general court and the specialized court. Then, the overall values for each court for each variable were compared. This allowed the study to find inconsistencies between the two courts. Then, further statistical analysis was performed to determine whether there was a statistically significant correlation between the outcomes of the cases and any other variables. The purpose of the further analysis was to determine which variables may have been causing the differences in case outcomes between the two courts.

In a second study, theoretical mathematical models were constructed for statistical analysis of judges' decision-making. The study investigated the effect of the incentive of reelection on judges' decision-making (Cohen et al., 2015). The study was interested in whether judges' sentencing decisions are inconsistent with their prior decisions in the time approaching an election where they are up for reelection. So, the study examined the consistency of each judge's decisions over time rather than consistency across judges.

In this study, a theoretical model was constructed to measure this type of consistency. The model was designed to consider the prior strictness or leniency of the judge's sentencing decisions as well as the incentive for the judge to make decisions that align with the public's opinions. The model was constructed to take data about a particular judge's past cases as input variables and return information about the level of change in the judge's decisions close to election time as output variables. Then, empirical data from real court cases were fed into the model. The outputs of the model quantified the significance of a change in judges' decisions before elections in which the judges were up for reelection.

In a third study, a statistical model was developed to test how certain factors affect consistency in individual supreme court judges' decision-making (Collins, 2008). This study used a heteroskedastic statistical model (a model that expects variation in the standard error of the data over time) to assess factors that might affect judges' decision-making, including the judge's political stances, the type of case, and the informational environment surrounding the case. Like the previous study, this one was concerned with consistency within individual judges' decisions, not consistency across decisions made by many judges.

This study constructed a statistical model to take data from court cases as input and return as output numerical values indicating how consistent the judges' decisions were across those cases. The model was heteroskedastic to account for expected fluctuations in judges' decisions due to their liberal or conservative political affiliations. The judges' decisions in cases from 1946 to 1995 were analyzed using the model. The outputs of the model were values for several variables related to the judges' consistency over time. Each variable had an expected range. A value within the expected range indicated expected trends for consistency of the judges' decisions based on prior literature. A value outside the expected range indicated that the factor represented by the variable was more inconsistent than expected.

The second consistency measurement method used when studying judges' decision-making is surveys. The surveys are designed to collect information from judges to determine which factors affect their judicial decision-making. Unlike statistical analysis, this method tends to be used to examine judges' decision-making processes rather than just decision outcomes. In some cases, the survey method is used to directly identify inconsistencies in the judges' decision-making processes, while in other cases it is used more broadly to find factors that affect the judges' decision-making processes. Two unique examples of this method of consistency measurement as applied to judges' decision-making will be discussed here.

The first example of the survey method is the use of decision-making scenarios combined with characteristic questionnaires. This method was used in a study that focused on the relationship between judges' characteristics and differences in their decisions (Van Koppen & Kate, 1984). First, the study simulated judicial decision-making by presenting many judges with the same set of written judicial decisional scenarios and asking them to make decisions about those scenarios.

The scenarios were constructed to simulate as closely as possible the content and structure of real court cases, except for increased brevity compared to real court cases. The judges were asked to indicate what ruling they would make on the case in the scenario on an eleven-point scale, with one end of the scale representing a very certain decision in favor of the plaintiff and the other end representing a very certain decision in favor of the defendant. The responses to the scenarios were coded based on several factors of interest, such as whether the judge's decision was for or against the more socio-economically powerful party, or whether the judge's decision was more aligned with formal law or the parties' specific interests. The judges' responses to the scenarios were found to be significantly inconsistent across several of the factors of interest of the study.

Then, the study also had judges fill out questionnaires about certain personal characteristics that the study determined might affect their judicial decisions. The questionnaire asked about characteristics including the participant's inclination to take risks, attention to others' needs, and self-esteem. Finally, the responses to the scenarios and questionnaires were analyzed to determine which personal characteristics were correlated to which factors of interest in the judges' decisions on the judicial scenarios. There was found to be little correlation of significance between any personal characteristics and the judges' decisions in the constructed scenarios.

The second example of the survey method is a study that used a questionnaire to examine which factors judges consider most important when making decisions in child custody cases (Reidy et al., 1989). As is often the case, the questionnaire was constructed by experts in the topics of the survey. In this case, those experts were psychologists and judges familiar with child custody cases. The study distributed a questionnaire to judges asking about several aspects of child custody dispute cases. The questionnaire asked the judges to rate the importance they assign to eleven distinct types of information when deciding on child custody cases. The judges were provided with a Likert-type scale on which to rank the importance of each type of information. The questionnaire also asked judges to separately rate the relative importance of twenty-four variables related to the outcome of cases where joint versus single-parent custody is being considered.

The mean rating for each type of information across judges was calculated. The ratings for several types of information were then compared. This comparison allowed the study to determine which types of information are taken under the most significant consideration during the judges' decision-making processes. The judges' mean ratings were also compared to the opinions of mental health professionals. While the correctness of responses is not usually a consideration when measuring consistency, this comparison allowed for an evaluation of how closely the judges' answers matched experts' opinions, an approximation of correctness that can be useful when evaluating consistency.

The questionnaire in this study also had several questions that asked participants which parent they would grant custody to if, in a particular case, all factors but one was equal between the two parents. The one differing factor was specified in each question.

The judges' responses to these questions were compared to find any inconsistencies in what the judges would decide based solely on the one factor singled out in the question.

3. Refugee Status Decisions

One method of measuring inconsistency was used in a study investigating refugee status decisions at the Swedish Migration Agency. The study aimed to test whether repeated exposure to emotionally demanding refugee status decisions would cause inconsistency in the workers' decisions (Pärnamets, Tagesson, & Wallin, 2020, p. 570). The methodology of this study included conducting a survey of decision-makers and caseworkers at the Swedish Migration Agency. The survey presented several fictional scenarios to participants. Each scenario described a potential applicant for refugee status and asked participants to rate how likely they would be to grant refugee status to the person in the scenario (Pärnamets et al., 2020, pp. 571-572). Three of the scenarios, one at the beginning, one in the middle, and one at the end of the survey, were carefully designed so that all relevant details were equal. Other details were changed to prevent the participant from recognizing them as the same (Pärnamets et al., 2020, pp. 571-572).

The study compared the responses to those three questions to determine whether there was a change in how likely participants said they would be to grant refugee status after being exposed to more scenarios (Pärnamets et al., 2020, pp. 571-572). By presenting fictional decision-making scenarios with only a specific factor of interest changing between scenarios, the researchers determined whether the variable in question, repeated exposure, influenced the consistency of the participants' decisions.

Another method of measuring consistency was used in a working paper which investigated the levels of inconsistency between decision-makers in Canada's refugee decision appeal process (Norris, 2019). The working paper used statistical analysis of refugee appeal decisions to quantify the level of inconsistency in the decisions. The study focused on decisions made by federal judges in cases where an applicant was initially denied refugee status and then appealed their case. In the Canadian refugee appeal system, the case is seen first by a first-round judge and then by a second-round judge. The study was interested in the level of consistency between only first-round judges as well as across both first- and second-round judges.

The study used several statistical equations to bound the disagreement between judges on refugee appeal cases. The statistical model analyzed data about past cases to determine the percent chance that any two judges within a specified group would disagree on any given case. This was done for two groups. First, the chance was calculated for first-round judges only. Then, the group was expanded to include both first- and second-round judges, and the chance of disagreement was calculated again. The average bound on a disagreement between judges was determined to be 16.9%, which was approximately halfway between the best-case and worst-case scenario estimates.

The study was able to quantify the level of inconsistency in the refugee appeal judges' decisions, however, it did not identify any potential causes of the inconsistency.

4. Human Resources

Most companies have a Human Resources (HR) department whose focus is to plan, coordinate, and direct the administrative functions of an organization. They oversee the recruiting, interviewing, and hiring of new staff, consult with top executives on strategic planning and serve as a link between an organization's management and its employees. Within HR, many decisions must be made, and it is important to keep those decisions consistent, so employees feel that they are being treated fairly.

One method used by some HR departments to measure the consistency of their decisions is called the "Full Consistency Method" (FUCOM). FUCOM is a comparison-based multi-criteria decision analysis procedure applying the principles of pairwise comparison and deviation from maximum consistency (Stevic & Brkovic, 2020). In other words, FUCOM compares decisions and determines which ones were successful based on the criteria inputted by the user.

A related consistency measurement method was used to investigate the accreditation process of the Accreditation Board for Engineering and Technology (ABET) (Ahmad, 2020). The objective of the study was to compile the critical success factors (CSFs) essential to achieve and sustain academic quality assurance and ABET accreditation in a systematic manner. Further, the research also attempted to identify the relative importance of CSFs using the fuzzy analytical hierarchy process (Fuzzy AHP) and full consistency method (FUCOM) with the help of decision-makers. Fuzzy AHP is the integration between Fuzzy sets theory and AHP to consider uncertainty involved in a decision-making process. The study used data analytics to measure consistency across the accreditation process.

5. Insurance

There are two major sources of inconsistency in the insurance field: imprecise ambiguity and conflict ambiguity. Imprecise ambiguity is a lack of precision in the data; for example, two insurers may have the same expected range that they believe a value will fall between, however the fact that it is a range and not just a point creates ambiguity. Conflict ambiguity is when two insurers each have a precise value, and the two are not in accordance with one another. In a study, it was found that insurers charge more when there is ambiguity because greater ambiguity means there is a greater risk, so there is less certainty about how much money the insurance company will have to pay, so the company charges more to cover a larger range of costs (Cabantous et al., 2011). The study also found that insurers' pricing decisions depended on the type of accident; they charge more for conflict ambiguity than for imprecise ambiguity in the case of flood and hurricane policies, but less for conflict ambiguity than imprecise ambiguity in the case of fire. The study used survey questions to measure knowledge consistency among insurers to determine inconsistencies in insurers' decision-making when there are ambiguities present.

Another study examined the effectiveness of the Basel II Accord within businesses in the United Kingdom to inform a survey of a major insurance company in the country (Bryce et al., 2016). Findings from the business field were used to analyze risk reporting and escalation via a survey that utilized structural equation modeling. It was found that attitude and uncertainty had major effects on an individual's intention to escalate operational risk. The analysis of survey responses included average variance and standard error, which gives a better understanding of the significance of the responses. The study found that attitude and uncertainty have a noticeable impact on the decision-making process. Even if someone has a high intention to report risk events and losses, if they see reporting these events as ineffective, harmful, or foolish, then they are less likely to report.

At insurance companies, measurement of the consistency of employees' performance is also necessary to determine the yearly bonus that each employee should receive. Employees who help the company earn the same amount of money should receive the same bonus. However, it can be difficult to measure how much each employee earns for the company each year because prepaid insurance policies earn all their money when they are sold, but a claim could be made at any time in their term, and this might lead to a loss. Many non-insurance companies use a bonus bank concept, in which their bonuses are paid incrementally if they have continued success or decreased if they do not. Another less popular method is that of a negative bonus, in which losses are carried forward. This is not ideal as it can lead some employees to only be motivated to meet their goals in productivity rather than surpass them. Bonus banks, however, do not lead to this problem and could be useful if they were applied in the insurance industry (Calandro, 2006).

Accident year analysis is another method that has only previously been used to measure insurance company performance, however, it could also be used to determine bonus payments for individual executives and employees. Insurance executives need to be consistent in their evaluation of their employees' success to give them fair bonuses, and there is a compensation approach that combines accident year analysis with the bonus bank and Insurance Performance Measure (IPM) to effectively align with the economics of the insurance business. This method can provide more consistency to insurance companies as it allows them to various bonuses depending on risk levels and size of policies, as well as resulting earnings over future periods (Calandro, 2006).

6. Healthcare

The health industry has many areas where measurements can be subjective and measurement of the inconsistencies between data collectors is necessary. Without consistency between technologists in clinical laboratories, there cannot be high levels of confidence in any study's accuracy. Inter-rater reliability, measuring how consistent data collection is between multiple collectors, and intra-rater reliability, measuring how consistent one collector is over time, are important measures of consistency in any medical study. While one might expect that intra-rater reliability would be high, it was found in one study that intra-rater reliability coefficients ranged from 0.15 to 0.90.

There are high levels of both inter-and intra-rater reliability for variables that have only two states, especially if they are highly differentiated. However, when finer discriminations must be made, there is more difficulty in obtaining reliability (McHugh, 2012).

Percent agreement is one statistic that can measure inter-rater consistency. This statistic is calculated by entering answers for multiple variables from two raters into a table and calculating the difference between the answers for each variable. For a simple calculation when there are two raters and only two values for each variable, the number of zeroes in the difference column is counted and divided by the total number of variables to get the percent agreement. Adjustments can be made when there are more than two raters or more than two values for each variable. For more than two raters, the percentage agreement of each row can be calculated, and the average of the rows can be found. More complex mathematical adjustments can also be made to account for the degree of difference in answers when there are more than two values for each variable (McHugh, 2012).

Cohen's kappa is another statistic that can be used to measure inter-rater consistency in healthcare. Cohen's kappa is measured on a scale from -1 to 1, with 0 representing the amount of agreement that is expected from random chance, and 1 representing perfect agreement. Negative kappa values are rare in practice, but when they do appear they represent disagreement or agreement worse than expected. There is a scale for values from 0 to 1.00. In terms of the agreement, 0.01–0.20 is none to slight, 0.21–0.40 is fair, 0.41– 0.60 is moderate, 0.61–0.80 is substantial, and 0.81–1.00 is almost perfect agreement. However, in the medical field specifically, 61% should not be considered substantial. Medical decisions must be especially precise, so 81% or greater should be considered a target value. Kappa is not a perfect statistic. Even if the kappa value is above ninety percent, this only indicates that 82-100% of the data is reliable (McHugh, 2012).

The Pearson coefficient is another statistic used for measuring consistency in healthcare. Some assumptions must be made to use the Pearson coefficient. These are debated, but they must include: (1) the data are derived from a random, or at least representative, sample, and (2) both variables are continuous, jointly normally distributed, random variables, following a bivariate normal distribution. Variables follow a bivariate normal distribution if both variables are normally distributed and if there is any relationship between the two variables, it is linear. To meet these guidelines, variables can be transformed to make them more normal. The Pearson correlation ignores outliers and treats each x-y pair independently. The Spearman coefficient has a range from -1 to 1, in which 0 represents no relationship and both 1 and -1 represent perfect monotonic relationships (Schober et al., 2018).

A final statistic used to measure inconsistency in healthcare is the Spearman coefficient. The Spearman coefficient has the same range as the Pearson coefficient, but unlike the Pearson coefficient, the Spearman correlation can be used to analyze nonlinear monotonic relationships, and it does not require a normal distribution. A Spearman coefficient is a Pearson correlation coefficient calculated with the ranks of the values of each of the two variables instead of their actual values. Both coefficients face some of the same problems as Cohen's kappa, in that

the interpretation of the coefficient is ambiguous. While many would agree that a coefficient that is less than 0.1 denotes a negligible relationship and those greater than 0.9 represent a strong relationship, there is disagreement over whether 0.65 should be considered a "good" or "moderate" correlation. Researchers in different fields can decide how strict they want the scale to be. In the case of medicine, they should err on the side of caution and look for greater coefficients (Schober et al., 2018).

In using these coefficients and formulas, it is important to remember that correlation does not imply causation. There could be some third variable that affects the two that are being compared, or it could be a mere coincidence that they are following similar trends. The agreement is also not guaranteed by correlation. Two variables can have a strong correlation but disagree, like in the case of one technique that consistently measures higher than another, which is an example of systematic error. Researchers are also advised not to rely on correlation coefficients, but to graph their data to allow for a visual inspection, as this can sometimes give a better idea of correlation than raw numbers (Schober et al., 2018).

Another place where consistency is critical in healthcare is in diagnosis. Differential diagnosis is a method of diagnosis that aims to differentiate ailments, especially those with similar symptoms. It begins with a list of conditions or diseases that the patient could have based on the patient's symptoms, medical history, lab results, and physical examinations. Once this primary diagnosis is developed, the doctor must answer a series of questions to eliminate ailments from the list, eventually reaching just one diagnosis, which can then be treated. The questions are asked in an order that will eliminate options with each step, and later questions are based on the answers to the previous ones. Differential diagnostic trees can be developed from these questions, as some questions will only be asked if a previous question is answered in a certain way. The inter-rater consistency between doctors' diagnosis processes can be measured by comparing the questions that they ask their patients during the diagnosis process and the order in which they ask them (Watt, 1997).

7. Loan Lending

Commercial loan officers evaluate a company's financial strengths and weaknesses. To effectively analyze these companies, loan officers must understand what is known as the "primary Cs of lending" which includes credit, collateral, capital, capacity, and character (Duchessi et al., 1988). A system is known as the Commercial Loan Analysis Support System (CLASS) is designed to evaluate where a company may have financial weaknesses and decide whether to approve the loan. The use of this system is intended to increase consistency in the loan officers' decisions. The loan officers look at factors such as the company's history of re-paying other obligations, their performance against the industry norms, the value of the company's assets, strength of leverage position, and if the loan can be supported by the company. CLASS offers a way for the officers to "easily enter and view financial data, choose execution options through menus, respond to queries, and obtain reports." (Duchessi et al., 1988).

When a company applies for a loan, up to five years' worth of their stored and recorded financial information is considered. CLASS can perform a trend analysis of the major financial trends for each company and provide the loan officer with the compiled information to evaluate whether it is improving, stable, or deteriorating (Duchessi et al., 1988). It also evaluates the credit, collateral, capital, and capacity of a firm by looking at the categories making up each of the Cs and giving an overall rating of poor, weak, normal, or strong in each category. CLASS will work for commercial loans of any size from many industries. The main limitations that come with this system are that it is not currently usable in commercial loans that have unique financial characteristics, and it is still the responsibility of the loan officer to look at the analysis and address factors that CLASS does not consider, which may introduce consistency.

One study assessed the consistency of commercial loans across various lenders using another specialized software (Carey, 2001). “This paper presents evidence about the consistency of rating assignments across lenders using Loan Pricing Corporation’s Loan Loss Database (LPC LLD). The data drawn from the LLD has information about individual commercial loans from the portfolios of more than two dozen lenders (all are banks) during 1994-98, including the internal rating assigned to each borrower, with such ratings mapped to a common ten grade LPC scale” (Carey, 2001). A “The least-preferred coworker scale” (LPC) is a management heuristic that assigns an individual's leadership style as either task-oriented or relationship-oriented. This study was conducted by using LPC with multiple lenders to map its ratings to a common ten-grade scale. The mapped grades are then recorded in the LLD. Mapping is an effective way of keeping consistency across different inputs that use different metrics to measure consistency.

8. Engineering

A study was conducted to measure process consistency between engineering consultants (Edwards et al., 2007). This study used a tool known as process maps. Process maps are useful for measuring process consistency. They provide a diagram of the steps in a process, allowing users to map out how they complete the necessary functions. These diagrams are provided to employees, with relevant steps of the process shown in the outer circles of the diagram. The employees draw arrows from step to step, numbering them to show the order in which they complete the process (see Figure 1 for an example of a completed process map). Finally, the process maps drawn by different employees are compared to determine the consistency between them.

The study also used Likert questionnaires to measure knowledge consistency. These consist of a sentence concerning the knowledge to be measured and a five-option scale from “Agree” to “Disagree.” It is essential to look at how often respondents provide the same answer when using these questionaries to determine consistency (Edwards et al., 2007). When respondents do not provide the same answer, it is important to note how significantly their responses differ. This method allows the posing of questions regarding different areas of knowledge to find which areas are consistent and which are not.

Requirement engineering is an essential piece of software engineering that also relies on consistency. For requirements engineering to be effective, the requirements must be consistent with given objectives and constraints. Certain metrics can be used to evaluate requirements based on the objectives and constraints provided. One paper includes six relationship metrics and five consistency metrics (Byun et al., 2014). Some of the consistency metrics relate to one objective and multiple constraints. Others relate to one objective and one constraint, multiple objectives and multiple constraints, or multiple objectives and one constraint. The final metric uses the value and cost of a requirement to measure its consistency. These metrics provide a rating of how much each requirement affects the given objectives or requirements.

The relationship metrics are similar, with the three main metrics corresponding to some of the consistency metrics. Degree of Constraints Conformance measures the relationship between one requirement and various constraints, Degree of Constraints Impact measures the relationship between various requirements and one constraint and Cost Demandable for each Requirement measures the cost of each requirement by considering the relationship among constraints. The other three metrics correspond closely to these three, but with a focus on the degree of these relationships. These include Degree of Objective Contribution, Degree of Objective Satisfaction, and Value Obtainable for each Requirement. While these relationship metrics present relevant degrees of a requirement with either objectives or constraints, the consistency metrics can present a relevant degree of a requirement with both objectives and constraints (Byun et al., 2014).

APPENDIX B

Interview Script for Senior Leaders

Introduction:

Hello, I'm Maggie Munroe. I'm Gabe Comenzo. I'm Harmoni Larrabee. I'm Rebecca Noris.

We are a student team from Worcester Polytechnic Institute. We are working with the Patent and Trademark Office for our junior year project. The USPTO has asked us to create a catalog of methods for measuring inconsistency in patent examiners' decision-making. To do this, we are gathering experts' opinions on factors that may cause inconsistency as well as which measurement methods are most effective and feasible to implement.

If you are willing, we would like to ask you a few questions about this topic. Gabe and I will ask questions while Harmoni and Rebecca take notes. Your answers will be kept entirely confidential and anonymous. You may choose not to answer any of the questions we ask, and you may request to stop participating in this interview at any time.

Are you comfortable with us recording this interview? We will be the only people to listen to the recording for the purpose of taking additional notes, and then the recording will be destroyed. For the record, we will ask you again once we start recording. (Make sure we get consent from all participants.)

(Start audio recording)

Are you comfortable with us recording this interview? (Make sure we get verbal consent from all participants.)

For our project, we are considering consistency in the patent examination process. Specifically, we are concerned with process consistency, knowledge consistency, and outcome consistency.

The goal of our project is to recommend specific methods for measuring inconsistencies within the patent examining process and identifying causes of those inconsistencies. We call these methods "consistency measurement methods."

Factors that cause inconsistency:

- What employee characteristics do you think might contribute to decision-making inconsistency between patent examiners?
- What knowledge discrepancies do you think might contribute to decision-making inconsistencies between patent examiners?
- What factors inherent in the patent examining process do you think leave room for inconsistencies in patent examiners' decision-making processes?
- Do you think there is more inconsistency in the evaluation of some types of patents than others? If so, which types of patents have more inconsistency?

Qualitative vs. Quantitative:

- Do you think qualitative methods for measuring consistency are useful? Why?
- Are there specific types of qualitative data that are most useful? If so, what are they?
- Do you think quantitative methods for measuring consistency are useful? Why?
- Are there specific types of quantitative data that are most useful? If so, what are they?
- Do you think either qualitative or quantitative data is more useful than the other? If so, which one and why?

Specific Consistency Measurement Methods (CMMs):

- For each CMM in the list below:
- Interviewer briefly describes the CMM.
- How well do you think the CMM aligns with the specific consistency-related factors that you are interested in measuring?
- How feasible do you think it is to implement this CMM the USPTO?
- Have you used this CMM before? If so, how effective was it?
- Have you used any CMMs we did not mention?
- What were they?
- How effective were they?

List of CMMs:

- Qualitative CMMs:
  - Likert Scales
  - Process Maps
  - Decision-Making Surveys
  - Questionnaires
  - Benefit-Risk Framework
- Quantitative CMMs:
  - Percent agreement
  - Cohen/Pearson/Spearman
  - FUCOM/MARCOS
  - Six Sigma Control Charts

APPENDIX C

Interview Script for Patent Examiner Supervisors

Introduction:

Hello, I'm [name].

We are a student team from Worcester Polytechnic Institute. We are working with the Patent and Trademark Office for our junior year project. The USPTO has asked us to create a catalog of methods for measuring inconsistency in patent examiners' decision-making. To do this, we are gathering experts' opinions on factors that may cause inconsistency as well as which measurement methods are most effective and feasible to implement.

If you are willing, we would like to ask you a few questions about this topic. [Name] will ask questions while [name] takes notes. Your answers will be kept entirely confidential and anonymous. You may choose not to answer any of the questions we ask, and you may request to stop participating in this interview at any time.

Are you comfortable with us recording this interview? We will be the only people to listen to the recording for the purpose of taking additional notes, and then the recording will be destroyed. For the record, we will ask you again once we start recording. (Make sure we get consent from the participant.)

(Start audio recording)

Are you comfortable with us recording this interview? (Make sure we get verbal consent from the participant.)

For our project, we are considering consistency in the patent examination process. Specifically, we are concerned with process consistency, knowledge consistency, and outcome consistency.

The goal of our project is to recommend specific methods for measuring inconsistencies within the patent examining process and identifying causes of those inconsistencies. We call these methods "consistency measurement methods."

Factors that affect consistency:

In your past experience as an examiner and your present experience as a supervisor:

- What characteristics of examiners contribute to decision-making inconsistency between patent examiners?
- What knowledge discrepancies contribute to decision-making inconsistencies between patent examiners?
- What factors inherent in the patent examining process leave room for inconsistencies in patent examiners' decision-making processes?
- Do you see more inconsistency in the evaluation of some types of patents than others? If so, which types of patents have more inconsistency?

Where consistency measurements can be applied to improve consistency:

Again, in your past experience as an examiner and your present experience as a supervisor:

- What kinds of consistency are most important to measure to improve the quality of patent examination?
- Where in the patent examining process is it most important to measure consistency?
    - Are there specific steps in the patent examining process where there is a high level of inconsistency? If so, what are they?
    - Are there any early steps in the patent examining process where inconsistencies have a significant impact on the rest of the process?
- How much time and effort do you think patent examiners are willing to put into participating in surveys or similar data collection about their examination process?

Specific CMMs:

- Did you use any strategies to improve your consistency as a patent examiner?
    - If so, which ones? How effective did you find them?
- Do you use any strategies to monitor or improve the consistency of the patent examiners you supervise?
    - If so, which ones? How effective do you find them?
- Do you think quantitative or qualitative methods for measuring consistency are more useful in your area of expertise?
- For each CMM:
    - How effective do you think this CMM would be for measuring patent examiner consistency? Why?
    - In which parts of the patent examining process do you think this CMM would be effective for measuring consistency?
- How would you rank the methods we have discussed?

APPENDIX D

Interview Script for Quality Assurance Personnel

Introduction:

Hello, I'm [name].

We are a student team from Worcester Polytechnic Institute. We are working with the Patent and Trademark Office for our junior year project. The USPTO has asked us to create a catalog of methods for measuring inconsistency in patent examiners' decision-making. To do this, we are gathering experts' opinions on factors that may cause inconsistency as well as which measurement methods are most effective and feasible to implement.

If you are willing, we would like to ask you a few questions about this topic. [Name] will ask questions while [name] takes notes. Your answers will be kept entirely confidential and anonymous. You may choose not to answer any of the questions we ask, and you may request to stop participating in this interview at any time.

Are you comfortable with us recording this interview? We will be the only people to listen to the recording for the purpose of taking additional notes, and then the recording will be destroyed. For the record, we will ask you again once we start recording. (Make sure we get consent from the participant.)

(Start audio recording)

Are you comfortable with us recording this interview? (Make sure we get verbal consent from the participant.)

For our project, we are considering consistency in the patent examination process. Specifically, we are concerned with process consistency, knowledge consistency, and outcome consistency.

The goal of our project is to recommend specific methods for measuring inconsistencies within the patent examining process and identifying the causes of those inconsistencies. We call these methods "consistency measurement methods."

Factors that affect consistency:

- What employee characteristics do you think might contribute to decision-making inconsistency between patent examiners?
- What knowledge discrepancies do you think might contribute to decision-making inconsistencies between patent examiners?
- What factors inherent in the patent examining process do you think to leave room for inconsistencies in patent examiners' decision-making processes?
- Do you think there is more inconsistency in the evaluation of some types of patients than others? If so, which types of patents have more inconsistency?

Where consistency measurement can be applied to improve the quality of patent examination:

- What kinds of consistency do you think are most important to measure to improve the quality of patent examination?
- Where in the patent examining process is it most important to measure consistency?
    - Are there specific steps in the patent examining process where there is a high level of inconsistency? If so, what are they?
    - Are there any early steps in the patent examining process where inconsistencies have a significant impact on the rest of the process?

Specific CMMs:

- For each CMM:
    - How well does this CMM align with the specific consistency-related factors that the USPTO wants to measure? ... with quality management?
    - How feasible do you think this CMM is to implement at the USPTO?
    - Have you used this CMM before? If so, did you find it to be effective?
    - In which parts of the patent examining process do you think this CMM would be effective for measuring consistency?
- Have you used any CMMs we did not mention?
- Do you think quantitative or qualitative methods for measuring consistency are more useful? Why?
- How would you rank the methods we have discussed?

APPENDIX E

Catalog of Consistency Measurement Methods

Data Collection Methods

Benefit-Risk Framework

General Description:

A benefit-risk framework is a method to weigh the benefits against the risks associated with a decision. To use this framework, participants think about and record all the benefits and risks they consider while making the decision. This allows for clear, consistent communication of the reasoning behind decisions to improve decision-making.

Implementation suggestions:

This method could be used in any area of the patent examination process that involves decision-making, such as the first office action, the patentability determination stage, or throughout the whole process. It could also be implemented at the management level where decisions are being made about the examination process.

During the patent examination process, examiners can use the benefit-risk framework to layout factors to find the best solution to a problem. For example, how to approach a patent or whether to keep searching for information. Supervisors can implement this framework to evaluate how effective different steps in the overall examination process are and the risks and benefits of how they currently work.

Rubric:

Factors	Type of Consistency	Applicability
Factors related to decision-making in the patent examination process such as an examiners' interpretation of training and problems, time management skills, and attention to detail.	Outcome consistency	Most applicable for decision-making throughout the patent examination process to make optimal decisions.
Accuracy	**Maturity**	**Other**
Participants may weigh benefits and risks differently or tend to take more risks than others, possibly reducing accuracy.	Applicable for assessment of both well-established and newly developed decision-making processes.	Advantage: Captures the evidence, uncertainties, and reasoning used to reach a decision. Disadvantage: The results of this method may be difficult to quantify.

References:

Kürzinger, M.-L., Douarin, L., Uzun, I., El-Haddad, C., Hurst, W., Juhaeri, J., & Tcherny-Lessenot, S. (2020). Structured benefit–risk evaluation for medicinal products: review of quantitative benefit–risk assessment findings in the literature. Therapeutic Advances in Drug Safety. https://doi.org/10.1177/2042098620976951

Brass, E. P., Lofstedt, R., & Renn, O. (2011). Improving the decision-making process for nonprescription drugs: a framework for benefit–risk assessment. Clinical Pharmacology & Therapeutics, 90(6), 791-803.

Decision-Making Surveys

General description:

In decision-making surveys, a series of fictional scenarios are presented to participants. The participant is asked how likely they would be to make a given decision in each scenario. A Likert-type scale is provided for the participant to give their answer. Sometimes, participants are asked to provide reasoning for their answers. Responses are collected anonymously and compared to find inconsistencies between different participants' answers to the same scenarios.

Implementation suggestions:

To implement this strategy for the measurement of consistency in patent examiners' decision-making, the fictional scenarios would be patent applications and the participants would be patent examiners. The scenarios might be fully fleshed-out fake applications, a specific section of a fake application, or an abbreviated description of a fake patent that includes necessary information but leaves out details that are not being considered as factors that are likely to affect decision-making consistency. In the case of difficulties in creating a fake patent, the surveys could potentially be run with a real application that has been processed in the past. If the past patent chosen is one that the patent office finds well-examined, then they can also compare the survey responses to the actual decision that was made in the past.

For each scenario, the examiners would first be asked the question "How likely would you be to approve the described patent?" The examiners would be given a Likert-type scale to give their answers. This scale might be a standard Likert with five or seven points ranging from "extremely likely" to extremely unlikely." However, scales that have more options, such as ranging from zero to one hundred, may provide more specificity in the data obtained from the survey. A study conducted by Pärnamets et al. in 2020 successfully implemented a 100-point scale in a survey of refugee status decision-making.

In addition to answering the above question on a sliding scale, it is also recommended to include a follow-up question for each scenario asking the examiner to describe the reasons for their answer. If a quantitative inconsistency is found in answers to the first question, then the reasons provided in the follow-up question provide qualitative information that may explain why inconsistency is present.

The fictional patents in the scenarios may be constructed to target specific factors of interest. If the variance of a factor within the patent application, such as the length of the application, is a factor of interest, the scenarios in the survey can be constructed to vary the factor of interest across the different scenarios but keep other relevant factors constant, to determine whether that factor is correlated with inconsistency in the examiners' answers.

The examiners participating in the survey may also be selected to target specific factors of interest. If a characteristic of the patent examiners, such as their level of experience, is a factor of interest, the survey can be given to examiners with a wide range of experience levels. Then, the examiners can be asked to report their level of experience in the survey and the results can be analyzed to determine whether the experience level of the examiner is correlated to inconsistency in responses to the scenarios.

Decision-making surveys will primarily be useful for measuring inter-rater reliability, as multiple examiners can look at the same patent and their decisions can be compared. However, this method could also be used to measure intra-rater reliability. To do this, an examiner would be given the same patent at different points in time and their responses would be compared.

Rubric:

Factors	Type of Consistency	Applicability
Factors related to the type or content of the patent being examined, such as length or complexity. Also, examiner traits such as years of experience.	Outcome consistency	Most applicable for the final decision of whether to approve a patent at the end of the patent examination process.
Accuracy	**Maturity**	**Other**
Participants may guess at what they believe the "correct" answer to be rather than providing honest answers, which reduces accuracy.	Applicable for assessment of both well-established and newly developed decision-making processes.	Advantage: This is the best method for gathering information on how different examiners respond to the same patent. There is a precedent of using this method at the USPTO. Disadvantage: This method is one of the most time-consuming, both in the creation of fictional scenarios and the time for participants to complete the survey.

References:

Pärnamets, P., Tagesson, A., Wallin, A. (2020). Inconsistencies in repeated refugee status decisions. Journal of Behavioral Decision Making, 33(5), 569-578. https://doi.org/10.1002/bdm.2176

van Koppen, P. J., & Kate, J. T. (1984). Individual differences in judicial behavior: Personal characteristics and private law decision-making. Law & Society Review, 18(2), 225–247. https://doi.org/10.2307/3053403

Questionnaires

General description:

Questionnaires consist of a series of questions, usually with multiple-choice options or a Likert scale to provide answers. Some questions may be open-ended and ask the participant to write their answer in sentences. The questions are compiled in a digital format and distributed to participants. Responses are collected anonymously and can be compared to find inconsistencies in participants' answers.

Implementation suggestions:

One way that questionnaires can be used within the USPTO to measure consistency is by measuring knowledge consistency between patent examiners. This could be done by giving patent examiners a questionnaire that asks them both multiple-choice and open-ended questions about the patent examination process. For the questionnaire to be effective, the questions must be developed carefully and without bias. The responses to the questionnaires can be compared to one another to find inconsistencies between patent examiners.

Another way questionnaires could be used within the USPTO to measure consistency is with patent examiners who have just completed their training. The questionnaires could be used as a "competency test" to make sure that the patent examiners have a consistent knowledge base that allows them to make consistent decisions. Likewise, the USPTO could use these questionnaires to make sure recently promoted Supervisory Patent Examiners have consistent knowledge.

Questionnaires require a plan and a well-designed script in order to be effective. If the script is not optimal, then it may be difficult to obtain useful data. Providing incentives to fill out questionnaires can also be useful because people are typically not inclined to fill out questionnaires unless they are provided with compensation for their time.

Rubric:

Factors	Type of Consistency	Applicability
The variables that questionaries can measure are extremely broad. Some of these variables include patent content, examiner characteristics, time of year, etc.	Knowledge consistency	Most applicable for testing knowledge consistency among examiners or SPEs at the USPTO.
Accuracy	**Maturity**	**Other**
Participants may guess at what they believe the "correct" answer to be rather than providing honest answers, which reduces accuracy.	Applicable for assessment of both well-established and newly developed decision-making processes.	Advantage: There is already infrastructure in place at the USPTO for distributing questionnaires, so they would be easy to implement. Disadvantage: Must be carefully designed and extensively tested to gather useful data. May receive poor engagement from employees

References:

Reidy, T. J., Silver, R. M., & Carlson, A. (1989). Child custody decisions: A survey of judges. Family Law Quarterly, 23(1), 75–87. http://www.jstor.org/stable/25739798

Process Maps

General description:

First, a diagram showing all the steps in a particular process is created. Then, the diagram is given to participants, who are asked to draw arrows to indicate the order in which they complete the steps of the process. Participants may indicate that they revisit steps multiple times or do not use certain steps at all. The diagrams are collected from participants anonymously and compared to find inconsistencies in the ways that different participants complete the process.

Implementation suggestions:

To implement this method, a diagram of all the steps in the process of interest must be created. To measure consistency in patent examination, the process would be the patent examination process as a whole or a particular section of the patent examination process, such as the early steps or how examiners put their findings into writing. For the most part, participants would be patent examiners. Supervisory Patent Examiners could be tasked with creating the process map templates for the examiners working under them, which must include all possible steps of the process. There is interest in using process maps in areas outside of examination as well, including quality assurance and application routing. In any case, the possible steps in the process must be agreed upon and laid out in a diagram, which examiners can then fill in by connecting the steps with arrows.

The completed diagrams are compared to one another to determine where and to what extent the examiners' processes differ. There is not a particular strategy for quantifying the inconsistencies in different examiners' process maps; this method lends itself primarily to qualitative evaluation. A study by Edwards et al. in 2007 is a good example of how process maps can be analyzed qualitatively to draw conclusions about process consistency among participants.

Rubric:

Factors	Type of Consistency	Applicability
Differences in patent examiners' examination processes. If used in tandem with a questionnaire about the examiners' characteristics, experience, and/or knowledge, it can also measure how those factors affect process consistency.	Process consistency	This method is applicable to the entire patent examination process; however, it may be most useful to assess stages of examination where examiners have the most choice in how they complete the process. Quality assurance personnel also showed interest in this method for their area of work.
Accuracy	**Maturity**	**Other**
If the process map is filled out after the examiner has already completed the examination process, memory bias may reduce accuracy. To increase accuracy, examiners could be asked to document their examination process as they go through it.	This method works best for well-established processes where the steps that may be taken through the process are already known and understood.	Advantage: This is the best method to find process inconsistencies. Disadvantage: The results of this method are more difficult to quantify than other methods. There is ambiguity around their development (what should be included, how specific the steps should be, etc.)

References:

Edwards, K., Jensen, K. L., Haug, A. (2007). Measuring process and knowledge consistency: A necessary step before implementing configuration systems. Innovative Processes and Products for Mass Customization, 77-88. GITO.

Data Analysis Methods:

Percent Agreement

General description:

This is a method of determining the level of agreement between different responses to the same question. There is a basic version of this method which can be used when a question only has two possible answers, but there are also more complex versions of the calculation which can be used for questions with multiple choice or Likert scale responses. The basic form of the calculation for an individual question involves counting the number of people whose responses agree and dividing that number by the total number of people.

Implementation suggestions:

Percent agreement can be used to analyze the results from process maps and Likert scales. This data analysis method can be used to measure the level of disagreement between the decisions of two or more examiners. It provides a percentage of how much agreement there is between responses to questions in a survey or questionnaire.

Rubric:

Applicability	Accuracy	Other
It can be used to quantify the agreement between two or more examiners' responses to a question.	It is very accurate for questions with only two possible answers. It can also be used when there are more than two answers, however it may less accurately represent trends in the data for this type of question.	It is easily calculated and directly interpretable, however it does not consider the possibility that participants may guess some answers, which may lead to overestimating the true agreement among participants.

References:

McHugh M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276– 282.

Cohen's Kappa

General description:

Cohen's kappa coefficient is similar to percent agreement; however, it factors in that sometimes when two people give the same answer, it's because they were both guessing and coincidentally chose the same answer, in which case they are not actually in agreement. It provides a value ranging from 0 to 1 that represents the correlation between the variables.

Implementation suggestions:

Cohen's kappa could be used to quantify the results of questionnaires or decision-making surveys. It can only measure agreement between two raters, but it is a useful base with related formulas that allow more than two sets of data to be compared.

This method could also be used to compare approval rates. While this comparison would not offer much information about the causes of inconsistency, it could be a first step that allows the USPTO to identify outcome inconsistency. It could be used to compare the approval rates of two examiners in the same area of art.

To use this method, data analysts at the USPTO will have to decide what they consider a significant correlation coefficient. Many of the existing magnitude guidelines related to the coefficient are arbitrary, with some considering a correlation coefficient between 0.61 and 0.80 substantial, while others in fields that require more precision aim for coefficients greater than 0.81.

Rubric:

Applicability	Accuracy	Other
Can be used to quantify the results of any of the data collection methods that involve Likert scales.	More accurate than percent agreement as it takes the chance of coincidental agreement into account. More accurate than Scott's pi, as that statistic assumes that annotators have the same distribution of responses.	Its use is limited to the comparison of only two raters. To compare more than two, Fleiss' kappa can be used, which is based on Scott's pi. There is also debate about what constitutes a significant correlation coefficient.

References:

McHugh M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Pearson

General description:

The Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations. It serves as a normalized measurement of the covariance and provides a value ranging from –1 to 1.

Implementation suggestions:

This method of data analysis would be most useful for analyzing the data that the USPTO already has than data obtained through the above methods. As it measures linear correlation, it could be useful for measuring changes in characteristics over time. These could include approval rates or salaries, among others. The method would only give an idea of how close the data is to a linear relationship, so if it isn't linear, this may not be the best option. If there is a set of data that should be linear, this can be used to measure how close it is.

The Pearson coefficient could also be used to measure interrater reliability. Approval rates could be compared between examiners with different ages or lengths of experience. Comparing these rates to a linear expression would give a good idea of whether they are remaining consistent over time. Or, if there is a goal production rate that examiners are aiming for, this will show if they are moving in the right direction.

Rubric:

Applicability	Accuracy	Other
Could be used in conjunction with the USPTO's database, especially for comparison of approval rates of an examiner over time or between examiners.	The value can be misleading if outliers are present. Since it only applies to linear relationships, the data should be graphed prior to using this method to determine if it is linear enough to warrant this method.	If the distribution is heavy-tailed or the sample size is small, the estimate provided by the Pearson coefficient will be biased, and therefore less useful. The data must be jointly normally distributed, but with a large enough sample this can be assumed.

References:

Schober, Patrick, et al. (2018). Correlation coefficients: appropriate use and interpretation. Anesthesia & Analgesia, 126(5), 1763–1768.
. https://doi.org/10.1213/ane.0000000000002864

Spearman

General description:

The Spearman rank correlation coefficient can be used to measure monotonic relationships between two continuous or ordinal variables. It is more often used with ordinal variables. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of the two variables. The coefficient ranges from –1 to 1, with these endpoints representing perfect monotone functions.

Implementation suggestions:

It could be used in any situation where the Pearson coefficient is used. If it is known that the relationship is linear, Pearson will be better suited, but if there's a nonlinear relationship, Spearman can be used. To use it, the values in the two datasets must be ranked, and then the Pearson formula must be used on the ranked values. Examiners could be ranked by experience and then Spearman could be used to compare their approval rates.

Rubric:

Applicability	Accuracy	Other
Could be used in conjunction with the USPTO's database, especially for comparison of approval rates of an examiner over time or between examiners.	It is accurate when used with monotonic relationships. If a relationship is known to be linear, Pearson could be more useful, but when it may not be linear, Spearman is a viable option.	Since the relationship only needs to be monotonic, it can be used with more data sets than the Pearson coefficient. If the data is not monotonic, the coefficient will be close to zero, but there could still be a significant relationship.

References:

Schober, Patrick, et al. (2018). Correlation coefficients: appropriate use and interpretation. Anesthesia & Analgesia, 126(5), 1763–1768. https://doi.org/10.1213/ane.0000000000002864

Six Sigma

General description:

A graphical tracking of a process input or output over time. In the control chart, these tracked measurements are visually compared to decision limits calculated from probabilities of the actual process performance. Control charts help show whether a process is in control and visualize exactly where the process needs to be worked on to decrease variability.

Implementation suggestions:

Six sigma control charts can be used to analyze data collected from decision-making surveys, questionnaires, and benefit-risk assessments. Six sigma control charts help visualize and analyze wherein a process the most variation is coming from. By using control charts, employees at the USPTO would know where to put their resources and attention to help decrease variation, and ultimately increase employee satisfaction. This method requires some way to measure data numerically rather than categorically.

Rubric:

Applicability	Accuracy	Other
Can be used to quantify the results of any of the data collection methods that involve numerical values.	Uses 6 standard deviations away from mean to help accurately point out variation within process. Shows whether a process is "controlled" or not.	Shows where the process needs improvement as well as visualizations using control charts. Can only be used with numerical data and requires quantification of qualitative data in order to use.

References:

Rejikumar, G., Aswathy Asokan, A., & Sreedharan, V. R. (2020). Impact of data-driven decision-making in Lean Six Sigma: an empirical analysis. Total Quality Management & Business Excellence, 31(3-4), 279-296.

APPENDIX F

Ethics

This project was conducted under the advisement of Worcester Polytechnic Institute (WPI), specifically, as an Interactive Qualifying Project. All research subjects remained anonymous, and all responses remained confidential. Participants in interviews were allowed to opt out of any questions they did not wish to answer. The opinions shared in this report are our own and do not represent the opinions of WPI or the United States Patent and Trademark Office. This information is solely used for academic research.