Evaluating the Validity and Reliability of Textile and Paper Fracture Characteristics in Forensic Comparative Analysis

Zachary Bailey Andrews
zba0001@mix.wvu.edu

Evaluating the Validity and Reliability of Textile and Paper Fracture Characteristics in Forensic Comparative Analysis

Zachary B. Andrews

Thesis submitted
to the Eberly College of Arts and Sciences
at West Virginia University

in partial fulfillment of the requirement for the degree of

Master of Science in
Forensic and Investigative Science

Tatiana Trejos, PhD, Chair
Keith Morris, PhD
Aldo Romero, PhD
Robert Ramotowski, MS

Department of Forensic and Investigative Science

Morgantown, West Virginia
2022

Keywords: textiles, paper, postage stamps, physical fit, performance rates, trace evidence, forensic science

ABSTRACT

Evaluating the Validity and Reliability of Textile and Paper Fracture Characteristics in Forensic Comparative Analysis

Zachary B. Andrews

In a comparative forensic analysis, an examiner can report that a physical fit exists between two torn or separated items when they realign in a manner unlikely to be replicated. Due to the common belief that it is unlikely that two unrelated fractured objects would match with distinctive characteristics, a physical fit represents the highest degree of association between two items. Nonetheless, despite the probative value that this evidence could have to a trier of fact, few studies have demonstrated such assumptions' scientific validity and reliability. Moreover, there is a lack of consensus-based standard protocols for physical fit comparisons, making it difficult to demonstrate the basis for the features that constitute a "fit." Since these analyses rely entirely on human judgment, they are highly subjective, which could be problematic in the absence of harmonized examination and interpretation criteria protocols.

As a result, organizations like the National Institute of Justice and NIST-OSAC have identified the need for developing standardized methods and assessing potential error sources in this field. This research aims to address these gaps as applied to physical fits of textiles and paper. Here, standard criteria and prominent features for each material are defined to conduct physical fit examinations in a more reproducible manner. Additionally, a quantitative metric is used to quantify what constitutes a physical fit when conducting comparative analyses of textiles and paper, further increasing the validity and reliability of this methodology and providing a manner of assessing the weight of this evidence when presented in the courtroom.

The first aim of this research involved the development of an objective and systematic method of quantifying the similarity between fractured textile samples. This was done by identifying relevant macroscopic and microscopic characteristics in the comparative analysis of a fractured textile dataset. Additionally, factors that affect the suitability of certain types of textiles for physical fit analysis were evaluated. Finally, the systematic score metric was implemented to quantify and document the quality of a physical fit and estimate error rates.

The second objective of this study consisted of establishing the scientific foundations of individuality concerning the orientation of microfibers in fractured paper edges. In comparative analysis of paper, it is assumed that the microfibers deposited across the surface of paper are randomly oriented, a key feature for addressing the individuality of paper physical fits. However, this hypothesis has not been tested. This research evaluated the rarity and occurrence of microfiber alignments on fractured documents. It also quantified the comparative features of scissor-cut and hand-torn paper and the respective performance rates.

Finally, the comparative analysis of textile and paper physical fits was validated through ground truth datasets and inter-examiner and intra-examiner variability studies. A ground truth blind dataset of known fits and known non-fits was created for 700 textile samples with various fiber types, weave patterns, and separation methods. Also, a set of 260 paper items, including 100

stamps and 160 office paper samples, were examined. The paper specimens contained handwritten or printed entries on two paper types and were separated by scissor-cut or hand-torn methods.

This proposed research provides the criminal justice system with a valuable body of knowledge and a more objective and methodical assessment of the evidential value of physical fits of textiles, paper, and postage stamps.

ACKNOWLEDGEMENTS

I would first like to express my endless gratitude and thanks to my research advisor, Dr. Tatiana Trejos. Dr. Trejos' guidance and support over the past three years has been paramount in my success as a student. You have provided me with countless opportunities for growth as a student, a researcher, a teacher, and, most importantly, a person. I cannot thank you enough for everything you have done for me over the years.

I would also like to thank my committee members: Dr. Keith Morris, Dr. Aldo Romero, and Mr. Robert Ramotowski. You all have provided invaluable advice and insight into this project that helped it succeed.

Thank you to the present and past members of the Trejos Research Group at West Virginia University. Whether it was helping with sampling, offering advice, or sharing a cup of coffee, I truly would not be where I am without you. Specifically, I would like to thank Oriana Ovide, Meghan Prusinowski, Evie Nguyen, Courtney Vander Pyl, and Korina Menking-Hoggat for welcoming me into the group three years ago with open arms. You made me feel at home, and I can never thank you enough for that.

I would like to thank our collaborator, Dr. Cedric Neumann, for all of his contributions into the statistical analysis aspects of this project. It has been a joy working with you, and I look forward to our future collaborations.

I would like to acknowledge Mr. Tristan Bond, Mr. David Neff, Dr. Michael Norton, and Professor Robert O'Brien for providing me with opportunities in high school and college to find a passion for science, introducing me to scientific research, and helping me become the scientist I am today.

Oriana, thank you for everything. Whether we were laughing, crying, stressing, traveling, getting tattoos, or watching Harry Potter for the first time, I knew that you would always be by my side, and I would never eat lunch alone. Our friendship has been the most wonderful part of the past few years and is something I will always cherish. Who knew that the summer day we spent breaking a windshield three years ago would evolve into such an amazing friendship?

Thank you to the Cream Cheese Danishes. You know who you are. I will always treasure our weekly dinners and trivia games, our trips for pickles, doughnuts, and tea, and our other ridiculous adventures. Your advice, support, and friendship mean the world to me.

Mom and Dad, I want to thank you, and the rest of our family, for your overwhelming and endless love, support, and encouragement as I follow my dreams. Thank you for everything.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

Trace evidence, and forensic science in general, relies on Locard’s Exchange Principle. Edmond Locard is known as one of the founding fathers of forensic science, and his theory holds true for practically every discipline in forensic science. Locard’s Exchange Principle simply states that “every contact leaves a trace.” That is to say when a crime occurs, it is presumed that the perpetrator will leave behind some trace material upon contact with any object or individual. The opposite is also true. The object will leave some trace material on the perpetrator upon contact. This interaction between two surfaces is also known as transference, a central trace evidence principle. From an investigation standpoint, the transfer principle is extremely significant, as finding a trace material, such as a fiber or glass fragment, embedded in a suspect's clothing can link the suspect to the crime scene [1].

The significance of transferred evidence can either appreciate or depreciate in any given case due to elements such as transfer level and direction. For obvious reasons, the discovery of two-way transfer is more significant than one-way transfer. Two-way transfer occurs when both surfaces transfer trace materials to each other, whereas one-way transfer occurs when only one of the surfaces transfers trace material to the other. The transfer level, whether primary, secondary, tertiary, or so on, also affects the significance of the evidence. Consider a breaking-and-entering case where the perpetrator breaks a window so they may enter a home. Primary transfer occurs as a direct result of the perpetrator breaking the window. This may take the form of a glass fragment embedding itself in the perpetrator’s wool sweater. Secondary transfer may occur following that when the perpetrator flees the scene. Perhaps the perpetrator hails a taxi cab or drives a vehicle. The glass fragment may then be transferred from the sweater to the automobile seat. This chain can and will continue as the trace material is passed from object to object. As these transfer levels continue, the probative value of finding trace materials at deeper levels decreases. This is because there is little evidence that the materials were ever in direct contact at these deeper levels. For example, the taxi cab seat was not in direct contact with the now-broken window despite containing a fragment of glass from that window.

This idea of probative value is why the concept of a physical fit is so vitally important in forensic science. A physical fit, by definition, directly links two items as originating from the same object. This occurs when the characteristics and features of the edge of a material, such as a piece of tape, aligns with the complementary characteristics and features on the edge of a second object, such as a roll of tape, in a manner that is unlikely to be replicated. Because fit demonstrates the two items were once part of the same object, a physical fit is often referred to as the highest degree of association between two items.

The 2009 National Academy of Science Report and the 2016 President’s Council of Advisors on Science and Technology Report were both highly critical of the then-current state of forensic science, highlighting the need for validity and reliability studies in comparative analysis disciplines, in addition to the establishment of error rates in these fields[2,3]. Despite this, little research has been done to establish the reliability and validity of physical fit analysis. The National Institute of Justice (NIJ) Forensic Science Technology Working Group (TWG) and the Organization of Scientific Area Committees (OSAC) recently published a call for research into the development of quantitative methods and the evaluation of error rates to assess physical fit comparisons[4,5], which this study will address.

A recent survey from the newly formed NIST-OSAC Trace Materials Subcommittee Physical Fits Task Group[6] and a publication from our research group[7,8] identifies textiles, polymers, and paper as the items most frequently submitted to forensic laboratories for physical fit examinations. Therefore, these materials were selected as the central aim of this thesis.

Textiles are a frequently encountered form of evidence at most crime scenes due to their ubiquity in everyday life. Textiles are encountered in clothing, carpeting, linens, and more[9]. They can be significant in violent crimes, such as assaults or murders, when the level of contact between a suspect and perpetrator is high, and textiles such as clothing may be damaged in a struggle[10]. Damaged textiles could also be encountered in breaking-and-entering cases if a suspect tears their clothing on a broken window, for example. If the suspect is apprehended, a physical fit could be identified between the separated fragment recovered at the crime scene and the article of clothing seized from the suspect.

Physical fits are determined through the examination of the physical features and characteristics of the material. Therefore, it is vital to first understand the composition and the manufacturing process of textiles before discussing the assessment of physical fits of this material. On a very base level, textiles are comprised of a multitude of fibers. Fibers are the basic units of textiles, similar to how nucleotides are the basic units of DNA. Fibers can be either artificial or natural. Artificial fibers are produced from polymers, which are forced through a spinneret and extruded into long fibers. In contrast, natural fibers are harvested from animals or vegetables and must be cleaned and combed. In either case, the fibers are then spun, or twisted, into yarns. The direction of the twist may prove to be forensically relevant, as it can be used to differentiate between textiles[11].

Once fashioned, yarns can then be used to produce fabrics. Fabrics are made by weaving, knitting, or braiding, among several other techniques, yarns together in some pattern. This production method, whether knit, woven, or otherwise, greatly affects the fabric's morphology, determining its strength, elasticity, texture, appearance, and more. Yarns that run north-to-south are known as warp yarns, whereas yarns running in the east-west direction are known as weft yarns. Generally, there are two different types of weaves in modern fabrics: plain weave and patterned weave. In a plain weaved fabric, the weft yarns alternate traveling under or over the warp yarns. In a patterned weaved fabric, the weft yarns are threaded through the warp yarns in some type of pattern[11]. A plain-knit fabric is one that involves interlocking the yarns to create a sort of "chain" of yarns. This method is often used when the manufacturer wishes to include intricate designs or multiple colors.

Textile construction and chemical composition are crucial components of textile examinations. These factors greatly influence how a textile will separate under force. Because of this, some types of textiles may not be suitable for physical fit analysis if they extensively stretch or deform during separation. Suitability determination will be a crucial element of this study. Some synthetic, polymer-based fibers, such as polyester, may deform easily upon fracturing, creating large amounts of distortion at the comparison edge, making a physical fit determination difficult and perhaps less reliable.

The second material considered in this research is paper, another common item at crime scenes. Paper is particularly interesting to questioned document examiners, who analyze typed or handwritten documents to determine if they have been forged or otherwise altered. This can be done through visual, microscopic, or chemical examinations of the hand-writing, ink, or paper[12]. This type of evidence is especially relevant in crimes such as fraud, kidnappings, or extortion cases where torn, cut, or shredded items can be used to establish potential links between pieces. Postage stamps are another commonly encountered form of paper in forensic science. Stamps can be found in bombing cases, ransom notes, and other mail-related crimes. A stamp on a mailed item of interest could be matched through a physical fit comparison to a booklet or sheet found in the possession of a suspect.

Office paper is manufactured from wood, which is predominantly comprised of cellulose fibers held together by lignin. Lumber must first be debarked and chipped before it can be converted into pulp. There are two types of pulping processes: mechanical and chemical[13]. The chemical pulping process involves the use of solvents to extract cellulose fibers from the lignin holding them in place. There are two primary varieties of chemical pulping methods: the Kraft method and the sulphite method. The Kraft method is the more commonly used method due to its efficiency and the quality of its product. The Kraft method uses a basic solution and pressure to soften the wood chips, which are then ejected from the solution into a vessel. The force of the chips striking the vessel frees the fibers from the lignin, producing pulp. Alternatively, the sulphite method uses an acidic solution to destroy the lignin holding the cellulose fibers in place. The mechanical pulping process uses heat, pressure, and machinery to grind and crush the wood chips into pulp. Using the same amount of wood, mechanical pulping produces twice as much pulp as chemical pulping[14]. After the wood chips have been converted into pulp, the pulp must be washed with water to remove any impurities or unwanted chemical solvents. The pulp is also filtered to ensure only the target fiber size is used in the final pulp. Following this cleaning process, the pulp is bleached to brighten or whiten its appearance. Certain chemical bleaching processes also prevent yellowing in the finished product.

Once the pulp has been prepared, it is ready to be turned into paper. The pulp is first passed through a fourdrinier machine, which is used for wet-end papermaking, the first stage of the papermaking process. In this step, the pulp is evenly distributed onto a wire mesh, removing water through gravity and suction. This leaves a wet paper sheet that is fed through several rolls to press and dry the sheet. This section of the manufacturing process is thought to be responsible for the assumed random orientation of the cellulose fibers in the paper. Excess water is pulled through the mesh, leaving the fibers behind, which form the sheet. Bajpai notes that if the fibers were oriented in the same direction, the paper would have poor strength and quality[15]. The paper continues on in the machine, passing through several presses that squeeze the paper sheet, removing water and condensing the fibers.

The paper is then carried to the dry end manufacturing phase, which removes the remaining water in the sheet. The paper is heated as it passes around several smooth, circular drums, which evaporate the water in the paper. The paper is then passed through calendar rolls which smooth the paper, and the finished product exits the machine, where it is stored in large rolls. The paper on the rolls can then be cut into the desired size. Depending on the type of paper and the manufacturer, coatings can be applied to the paper for functional or aesthetic purposes.

1.1 LITERATURE REVIEW

Reports describing the use of fracture-matching in casework comprise the bulk of the literature surrounding physical fit analysis. A significant portion of this literature is devoted to metals, often taking the form of a toolmark examination. For instance, one case study describes the use of striations as a feature of interest for comparing the tip of a broken screwdriver found at the scene of a breaking-and-entering case with a broken screwdriver found in the suspect's vehicle. In another case[16], striations found in metal, in addition to edge morphology, were used to compare two halves of a broken knife used in a stabbing. Other studies relate physical fit examinations with firearms examinations, describing the comparison of bullet fragments to identify the number of shooters present at a crime scene[17] and the number of firearms used in a shooting[17], as well as the comparison of fractured elements of a single firearm used in a robbery[18].

One recent study[19] discussed physical fits of tool steel fragments and the reliability of casting materials, such as mikrosil, to adequately capture the necessary details to conduct an accurate physical fit examination. Practitioners widely use casting materials to make replicas of evidence, which is often used for comparative analysis, so it is crucial to determine if the casting material reproduces microscopic features in the detail required for these examinations. This study used confocal microscopy to create three-dimensional images of the samples, and it found that casting material could accurately reproduce the original evidence in great detail. Quadratic discriminant analysis was used to classify matrices for 300 combinations of original and replica pairs; zero misclassifications were reported. However, the authors did note that the casting material can be hindered by the examiner that creates the case, as uneven pressure of the evidence into the material can cause dissimilarities in the replica.

Other prominent materials studied include plastics and paint. One report describes the features of interest that were useful in reaching a physical fit conclusion for soft plastic bags, including garbage bags. Striations were again identified as a relevant feature, as were perforations, surface scratches, and others[20]. Manufacturing knowledge was critical in this study, as the orientation of specific markings on the garbage bags could be used to ascertain the production sequence[21]. Paint can be especially relevant in cases involving automobiles and identifying a physical fit between a paint chip and a vehicle can link that vehicle to a crime. Case reports have highlighted the importance of striation alignment when examining a potential physical fit for a paint chip[22], while others discuss the impressions that can be left in the paint by welding beads, door frames, and other objects[23,24].

Quantitative assessments of physical fits are also frequently discussed in the literature, especially regarding duct and electrical tapes. One study[25] using duct tapes addresses examiner performance by describing accuracy rates for four examiners who conducted physical fit comparisons for several sets of duct tapes with known ground truth. The sets consisted of both hand-torn and scissor-cut tape pairs. While there were no false positive or false negative misclassifications, some pairs were classified as "inconclusive." Additionally, a higher portion of scissor-cut pairs was classified as "inconclusive" (19%) compared to hand-torn pairs (8%). The same author conducted a separate study on electrical tape, using sets with variations in separation method, tape brand, and the person creating the tape pairs[26]. The ten total sets contained 106 known true fits. Of these 106 true fits, 96 were correctly classified, while seven were deemed inconclusive, and one was misclassified as a false positive.

Tulleners and Braun laid the foundation of the present study in their 2011 study that outlined a method for assessing the quality of a duct tape physical fit[27]. The authors proposed a "% Match" metric that is calculated by measuring the distance of the matching area along the comparison edge in centimeters and dividing it by the total length of the comparison edge. A "Match Category" is also described that allows the examiner to express his/her confidence in a given physical fit by assigning it one of six classifiers. A follow-up study discussed examiner performance for a set of 1600 duct tape pairs, containing both torn and cut tapes[28]. The authors detailed high accuracy rates and low misclassification rates for each set.

Prusinowski et al. built upon this work by defining a systematic approach for the quantitative assessment of duct tape physical fits [7]. In this study, the authors employ an Edge Similarity Score (ESS) to quantitatively define the quality of a physical fit between two pieces of duct tape. Duct tape is comprised of three general layers: the adhesive layer, the reinforcement layer, and the backing layer. The reinforcement layer, also known as scrim, is made of fibers oriented in both the warp and weft directions, forming a natural way to subdivide the length of a duct tape fracture into comparison areas. Each scrim bin represents a comparison area. The ESS is a ratio between the number of matching comparison areas and the total number of comparison areas. The authors used this ESS metric to assess 2280 duct tape comparisons and reported high overall accuracy and zero false positives. The study included several different qualities of duct tape that were fractured using both hand-torn and scissor-cut separation methods. It was found that ESS values higher than 80% were indicative of a match, while ESS values lower than 20% were indicative of a non-match.

A recent article from the Netherlands Forensic Institute[29] also discussed the quantitative assessment of physical fits of duct tapes, detailing the use of loopbreaking patterns to calculate likelihood ratios that assess the evidential value of the comparisons. Loopbreaking practices refer to the different ways that the loops created by weft-insert scrim fibers can fracture. This study considered 136 pieces of duct tape of three different qualities. Two tearing methods were also used: top-down and bottom-up. Top-down tearing is a tear created by ripping from the top of the duct tape and moving downward, whereas bottom-up tearing is done by starting at the bottom of the duct tape and pulling upward. The examiners compared the edges of torn duct tapes and determined if the damage they observed to the weft-insert loops corresponded between both samples in the comparison pair. A Bayesian network was created using these loopbreaking patterns to evaluate likelihood ratios. This method produced very strong evidence for either fits or non-fits for most comparisons, along with high accuracy.

To date, no such systematic, quantitative approaches exist for physical fit comparisons of textiles. One fractography study discussed the type of textile damage that can be observed after a stabbing event, detailing the differences in damage done by a dull blade compared to a sharpened blade and how fabric construction can impact the edge morphology of a fractured textile sample[30]. Another such study describes several more factors that can affect a fracture. In addition to the above elements, the authors also found that a serrated blade causes more distortion along the fracture edge than a straight blade. Interestingly, it was also found that fractures measured four hours after the stabbing event were found to be longer than when the same fractures were measured immediately after the event. This was thought to occur as a result of the yarns near the fracture site unraveling and loosening as time progressed.

Despite these fractography studies, no empirical studies have been published that formally assess performance rates for physical fits of textiles or describe specific features and characteristics that could be useful in a physical fit comparison. The previously discussed empirical studies on duct tapes[7,27,28] provide a strong foundation for work with textiles due to some common characteristics between textiles and tape, such as the use of a woven reinforcement layer in duct tape that is of similar construction as a traditional textile. Despite this similarity, there are significant differences in the physical and chemical structures of textiles and duct tapes that require developing and evaluating a fit-for-purpose method.

It is essential to recognize the current status of the textile examination discipline and how this research seeks to fill gaps and improve the general understanding of the field. Sloan et al. describes the current state of textile examination in their 2018 article by exploring how practitioners have responded to the 2016 PCAST Report, specifically how Australian agencies have established a Textile Damage Working Group (TDWG) to provide guidelines on best practices in forensic textile examinations [31]. One primary function of the TDWG is empirical testing to establish the validity and reliability of textile examinations, a vital component of this research, as well as a critique levied by the PCAST report against feature-based comparisons in general. The TDWG began collaborating with the Australian Federal Police in the first half of the last decade to establish inter-laboratory studies to better understand the reliability of these types of comparisons. These studies use ground truth casework examples to assess examiner proficiency and inter-laboratory variation in results. The authors also discuss the effect of cognitive bias in comparative analysis, describing a sequential unmasking strategy for revealing case information to examiners. Using this strategy, the examiner would only have access to this unnecessary information after documenting a preliminary examination of the textile. Often in these cases, examiners will attempt to determine the amount of force or type of weapon used in an assault or murder case based on observations made about the level of damage to the textile in question.

Textile evidence is often encountered in cases involving sharp force trauma. Due to the nature of these stabbing events, damaged textile evidence is usually recovered. Kemp et al. discusses this evidence in their 2009 article [30]. Of particular interest to this research is the effect of blade sharpness on textile damage. Generally, textile damage inflicted by a dull blade will cause some distortion at the fracture edge. This distortion takes the form of extraneous fibers as the yarns at the edge are in general disarray. Rather than severance, the fibers are separated via tension. As the fibers fail, the fracture occurs. On the other hand, textile damage inflicted with a sharpened blade will be much more orderly, lacking any extraneous fibers due to a cleaner separation method. Kemp and colleagues also noted the effect of fabric composition on edge morphology. It was determined that drill, or diagonally woven fabric, and knit fabric could be distinguished based on the damage that was done using the same kitchen knife. Another morphological change occurs at the fiber level. Fibers cut by a sharp object, such as a knife or scissors, are often cinched or pinched at the end. Fibers severed by blunt force often display a “mushroom-cap” at the fracture point.

Fiber degradation, generally due to laundering, also played a part in determining the edge morphology.

Cowper, et al. build upon Kemp's work by analyzing the various factors involved in a stabbing event with textiles [32]. Factors considered include blade type, fabric type, fabric degradation, and fabric extension. Regarding fracture length, it was found that fractures measured 4 hours after the stabbing event were significantly longer than fractures measured immediately after impact (mean fracture length immediately after event: 28.3 mm, mean fracture length four hours after event: 29.8 mm). This highlights the need for examiners to consider the length of time between their analysis and the fracture event when conducting their examinations. It was also found that male participants produce longer fractures than female participants. Blade types (kitchen knife or serrated bread knife were also found to affect fracture length in conjunction with fabric type. While the straight kitchen blade produced no significant difference in length for either type of fabric used in this study, the fractures created by the serrated bread knife did significantly differ in lengths, with the 100% cotton producing a longer fracture than a cotton/elastane hybrid. Only general observations could be made about morphological features due to inter-participant variability. It was found that the fracture edges caused by the serrated bread knife were more distorted than those created by the kitchen knife. This is to be expected, as the serration may act similarly to a blunter instrument, tearing and ripping the fibers rather than severing them. It is also noted that more distortion was seen on the cotton fabric rather than the cotton/elastane blend. This is interesting, as typically, polymer-based fabrics like elastane (otherwise known as spandex) produce more distortion than natural fabrics such as cotton.

The existing literature for physical fits of paper fragments is predominantly limited to algorithm-based document reconstruction. Studies by Ukovich & Ramponi, Lotus et al., and Kleber et al. all detail computer algorithms that can be used to reconstruct both hand-torn and shredded documents. Ukovich & Ramponi's work identifies three different features that are used throughout the reconstruction process [33]. These include color features, features that allow the algorithm to detect lined paper, and features that describe handwriting. A Hough transform was employed by the authors to detect squared patterns. This required that the shredded strips of notebook paper used in the study be digitally divided into squares. A discriminatory metric is then assigned to each fragment of shredded paper.

Kleber and colleagues, in 2009, discussed an algorithm that analyzes the rotational orientation of torn paper fragments as well as the color of both the paper itself and the ink on the paper to align torn fragments [34]. The algorithm assessed 678 fragments for rotational orientation and found that only 32 could not be properly oriented, which was attributed to lack of a straight edge or lack of graphical content on the fragment. The mean error for rotational analysis was found to be 1.95 degrees. Color segmentation was able to differentiate between color and black text. This algorithm shows promise for general use for fragment organization prior to document reconstruction.

In the work done by Lotus et. al. in 2016, hand-torn fragments of paper were digitally scanned [35]. A polyline simplification algorithm then extracted the contours of the edges of the torn fragments. The polygons generated from the contour edge were processed by analyzing the number of sudden changes in the orientation of the contour and the Euclidean distance between the vertices of the polygon. Comparisons between torn fragments were assigned high scores if the Euclidean distance was small and the number of sudden changes in contour orientation between sides was equal.

Ten years earlier, in 2006, Justino et. al. uses a similar algorithm as the aforementioned Lotus study, whereby a polyline simplification algorithm is used to reduce complexity in fragment shape [36]. After the contours of the fragment edges are extracted and transformed into polygons, two features are used for fragment comparisons. The first is the angle of the vertex of the polygon with respect to its neighbors and the second feature is the Euclidean distance between the vector and its neighbors. A match is declared if the vector angles between two fragments sum to 360 degrees and the Euclidean distance is similar. Reconstruction of an entire document is made possible by comparing a fragment with all other fragments. A match is then made between two fragments using the similarity metric and the two fragments are merged into a single fragment. This process is repeated until the entire document is reconstructed, essentially resulting in a single large fragment. The authors found that this method is best used in combination with a human examiner, and is by no means a replacement for one, as the algorithm's performance decreased when reconstructing documents with increasing number of fragments.

In 2008, Smet developed an interesting strategy for reconstructing torn documents that are recovered in a stack at the crime scene [37]. A "single page, left-most on top" ripping strategy is assumed, where the document consists of only one page and the tearer places the fragment(s) in his/her left hand on the top of the stack when tearing the document in half. It is also assumed that no fragments of the original document are lost, and no foreign fragments are inserted into the stack. For a stack comprised of four fragments, the algorithm would match the fragments in the first (top) and third positions and the fragments in the second and fourth (bottom) positions. The matched fragments would be merged into a single fragment, and fragment 1-3 and fragment 2-4 could be matched and merged into the final reconstructed document. The algorithm can also be adjusted if the tearer used a "right-most on top" ripping strategy. If the a "flip" strategy is used, where the tearer alternates left-most and right-most on top, then a more significant modification to the algorithm would have to be made. For N number of fragments, Smet found that the algorithm is generally N to 2N times faster than randomly searching for matching fragments. Unfortunately, Smet does not discuss any particular fragment features that are identified in the matching process.

Diem et. al.'s 2010 study of 690 document fragments identified several features that can be used in document reconstruction [38]. These features include the color of the paper and the text on the paper, the segmentation of the paper (lined, gridded, blank), the type of print (handwritten or typed), and how the document fragment can be rotated based on the alignment of the text on the paper, if present. This final feature, however, was found to fail when there was no content on the fragment, or, in other words, when the fragment was blank. The rotational analysis also struggled when the fragment did not have any straight edges. Identification of line segmentation was also problematic, resulting in a 15% false positive rate. Finally, this method confused typed fonts that were designed to appear hand-written.

In 2014, Li et. al. developed a feature-matching algorithm to reconstruct shredded documents [39]. This algorithm can consider both vertically and horizontally shredded documents, as well as front-and-back print. The algorithm operates by converting the pixels of the scanned fragments into binary. A similarity metric is calculated by dividing the number of matching pixels by the number of total pixels in a comparison. While no error rates are given, the speed of the algorithm is discussed. The authors specify that an 11 line by 19 row double-sided document takes only 10 milliseconds for the algorithm to reconstruct. Additionally, the algorithm was successful reconstructing documents with English or Chinese text.

Marilyn Aguilar's 2019 article *Physical Match: Uniqueness of Torn Paper* discusses the assumption of the uniqueness of a torn paper fragment edge [40]. In this study, a random naming scheme was applied to 50 index cards, which were torn in half. All of the halves were placed in a large bag and mixed, and six halves were pulled out of the bag and set aside. By removing 6 halves, a total of 44-47 potential matching pairs remained in the bag, depending on the ground truth of the six halves that were removed from the sample pool. The remaining 94 halves were removed from the bag and analyzed by an inexperienced examiner, who identified 44 matching pairs of fragments with zero errors. While the author did correctly identify every matching pair, they failed to describe any features along the fracture edge that contributed to their decision-making process.

In summary, the forensic community can benefit from research that can demonstrate the scientific foundations of physical fit examinations and fill up the significant gaps for physical fit comparisons of textiles, paper, and postage stamps. There are neither empirical studies assessing performance rates for physical fit comparisons of textiles, nor are there studies that identify relevant textile features that can be used in a physical fit comparison. Therefore, determinations of a physical fit in paper or textiles are left in great part to the judgement of the examiner, without standardized and validated protocols. This study seeks to address these shortcomings by assessing error rates for a large dataset of textiles of varying composition, construction, design, and separation method. It will also establish a basis and criteria for comparisons, determine the relevant features for examination and assess their effect on the quality of a physical fit, and develop protocols for documenting, reviewing, and reporting results. The literature regarding paper materials is predominantly focused on algorithm-based document reconstruction of shredded documents, and the assumption that microfibers are randomly disritbuted across the surface of paper has not been proven. Therefore, this research also seeks to develop and assess a human-based methodology for physical fit comparisons of paper materials and evaluate the use of microfiber alignment as an indicator of a physical fit for these materials.

2. OVERALL OBJECTIVES OF THE PROJECT

2.1 OBJECTIVE I

The continued development of the Edge Similarity Score metric for physical fit comparisons of textiles was focused on identifying relevant characteristics and defining the associated terminology and systematic documentation in the comparative analysis of fractured textile sets. The characteristics identified in this study include weave alignment, fiber gap alignment, pattern alignment, and continuation of fluorescence. These characteristics enhance the quality of a physical fit and increase the confidence of the examiner when he/she concludes that a comparison is a “fit”. Other characteristics, such as distortion, curling, and secondary tearing can diminish the quality of a physical fit and decrease an examiner’s confidence when making a conclusion in a physical fit comparison.

In addition to the aforementioned relevant characteristics, there are also other factors that may affect the suitability of this method’s application to textiles. For example, preliminary work indicates the composition of the textile can significantly affect error rates in fits examinations. It was observed that more easily deformable fabrics, like polyester, are subject to significant amounts of distortion when a fracture event occurs. This distortion severely impacts an examiner’s ability to accurately conduct a comparison. For these reasons, fabrics should be screened early in the process for their suitability for physical fit comparisons, otherwise, they can lead to unacceptably high error rates.

For this research, a large dataset of textile comparison pairs was created. The analysis of this dataset, which consists of a wide variety of sample types, provides an opportunity to continue documenting these relevant features and observing how they can vary and affect an edge comparison. Identification of these features and the application of the ESS method will help the broader scientific community gain a better understanding of what constitutes a physical fit in this type of analysis by demonstrating specific characteristics that are required to achieve a “fit”. Reliability of the ESS method was also tested through the comparative analysis of this dataset. Error rates, which are sorely lacking in this discipline, were established, increasing the veracity of these types of analysis and providing practitioners with scientifically-sound data that can support the examiners opinion in the courtroom.

A dataset of 700 total textile comparison pairs was created. Six hundred of the 700 pairs were split between knit and woven samples of various patterns and designs. These 600 samples were comprised of cotton. Furthermore, this set was split evenly to account for both hand-torn and stabbed samples. This set sought to address the effect of different factors, such as separation method and construction, on the quality of a physical fit. The remaining 100 pairs were comprised of various fabric types, such as polyester and rayon to further assess the suitability of the method and investigate intra- and inter-examiner variability A sampling diagram can be seen below. The dataset consists of both known fits and known non-fits. These samples were compared by examiners who were blind to the ground truth. Additionally, each comparison pair was imaged, and the features were documented using an Epson Expression 12000XL scanner.

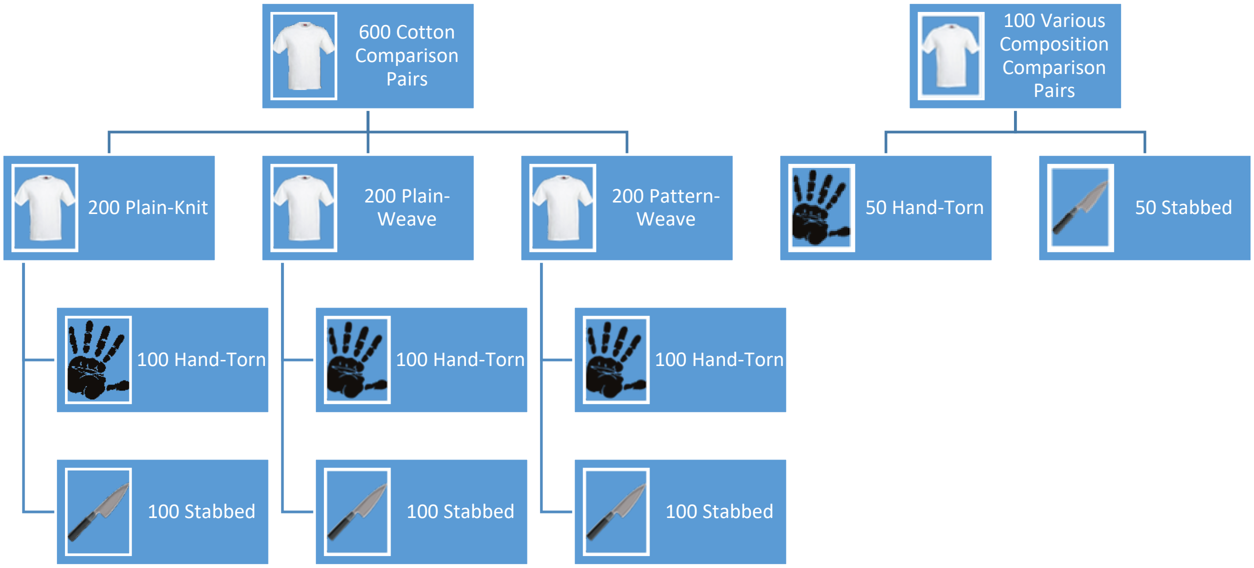


Figure 1. *Diagram detailing the makeup of the textile dataset, including the number of samples within each subset.*

Experimental design and results for Objective I are discussed further in Chapter 3. All data files for this chapter have been archived in a central database.

2.2 OBJECTIVE II

The second objective in this study involves the establishment of the scientific foundations of the individuality of the orientation of microfibers on paper fracture edges. This was done by evaluating experimentally the assumption that the random alignment and orientation of fibers in a piece of paper is rare and therefore, when present, it provides individuality along the fracture edge. It also includes quantifying the compared features on both hand-torn and scissor-cut paper and the performance rates on comparative analysis of papers from the differing manufacturers and ink entries (printing or hand-writing)

One area that was further explored in this research was the micro-features that contribute to a physical fit in addition to the macro-features that have been explored with tapes and textiles thus far [41,42]. Due to this, the ESS metric may not be sufficient for physical fits of paper in its current form. Rather, the comparative analysis of a paper physical fit will be more similar to the comparison of two fingerprints. Using this analogy, the alignment of microfibers across the comparison edge can be likened to the presence of identical minutiae points on two fingerprints. Therefore, the number of micro-features in agreement to achieve a "fit" were considered, and a classification threshold was investigated for classifications of unknown samples.

One key component of this research was analyzing known ground-truth samples. To prove that the microfibers on the paper surface are, indeed, rare on random matches, it is expected that there would be no alignment of microfibers along the comparison edge of a known non-fitting pair of samples. Likewise, it should be expected that this microfiber alignment would be found along the comparison edge of known fitting pairs. By analyzing a large dataset consisting of both known fitting and non-fitting pairs, the veracity of this assumption was assessed.

A dataset of 160 paper comparison pairs was created. Due to its general ubiquity, this study focused on white office/copy paper. The paper was sourced from different manufacturers in order to assess variability and some entries were added to include typed text or handwriting to simulate realistic casework. Two separation methods were employed, hand-torn and scissor-cut, and the dataset was split between fits and non-fits. All comparison pairs were documented using a Zarbecco MiScope portable digital microscope to generate images of the comparison edge of each pair. In Figure 2, a sample diagram for this dataset can be seen.

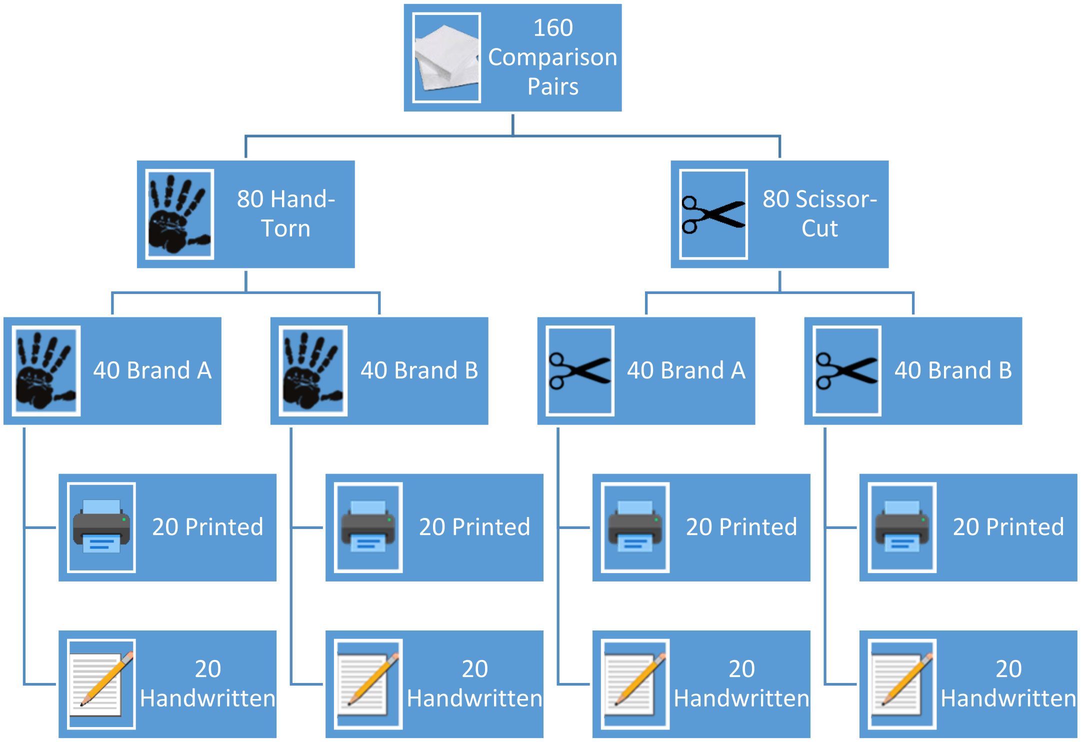


Figure 2. *Sample diagram highlighting the breakdown of the 160 comparison pairs of the paper dataset.*

Methods and results for Objective II are discussed further in Chapter 4. All data files for this chapter have been archived in a central database.

2.3 OBJECTIVE III

The third objective of this study builds upon the second objective by transitioning from paper to postage stamps. While postage stamps are also made of paper, there are several key differences that effect their ability to undergo physical fit analysis. When comparing two stamps for the presence of a physical fit, it often means comparing two completely whole postage stamps, and not a single stamp that has been fractured in half. This is because postage stamps are printed together on sheets of several stamps, typically 20, and the individual stamps are cut into the sheet by a machine. Therefore, physical fit comparisons of stamps are often cases determining if a stamp originated from a specific place on a sheet, to evaluate if the evidence supports the hypothesis of same-source as opposed to different-source.

As a result of the machine-cutting process, postage stamp edges are uniform. This means that the edge similarity score metric must be adapted in a unique way, as all stamp edges of the same stamp design are expected to fit together, regardless of ground truth, unless there is some kind of damage to suggest otherwise. This causes the focus of the comparison process to lay in the alignment of microfibers between the stamp edges. It has long been assumed that microfiber alignment would not be observed in non-fitting edges, though this research disproves that assumption. Microfiber alignment was, indeed, observed among true non-fitting stamp edges. As a result of this finding, the implementation of a classification threshold was investigated to determine if a given comparison should be classified as a fit or a non-fit based on the number of aligning microfibers observed across the edge.

A dataset of 100 postage stamp comparison pairs was created using three sheets of postage stamps. Each stamp was removed from its origin on the sheet and placed on an individual acetate square so that both the upper design side and lower adhesive side of the stamp could be viewed by an examiner. This process was documented to preserve the ground truth of the dataset. After 100 comparison pairs were created by a third party, two examiners began the analysis of the dataset independently and were unaware of the ground truth to minimize bias. The comparisons were conducted by subdividing the edges of the stamps into 13 bins of equal size and conducting bin-by-bin comparisons between a pair of stamps to identify any aligning microfibers. The location within each bin and a short description of each aligning microfiber observed by either examiner were documented. The initial 35 comparisons performed by the examiners were re-evaluated and discussed, and new terminology and criteria was further established to lead to better consensus between examiners. All comparison pairs will be documented using a Zarbecco MiScope portable digital microscope to generate images of the comparison edge of each pair.

Methods and results for Objective III are discussed further in Chapter 5. All data files for this chapter have been archived in a central database.

2.4 DATA ANALYSIS AND STATISTICAL ANALYSIS

This research will further the implementation and validation of the edge similarity score metric in the physical fit discipline. The edge similarity score, or ESS, is a quantitative metric used to measure the similarity of the edges of a fractured material. The metric defines the quality of the alignment between two materials, otherwise known as a physical fit, as the ratio between the comparison areas in agreement and the total number of comparison areas. This value is then reported as a percentage. This metric allows an examiner to give a numerical weight to their conclusions in reports and in the courtroom. Rather than simply reporting an overall match between a known and questioned sample, an examiner can report that the edges of the known and questioned samples had a quality score of 90%, for example. Because this metric breaks down the comparison process into small comparison areas along the edge, the examiner can also precisely document which characteristic features they have identified, the exact location of these features, and their opinions regarding these features. This detailed documentation helps strengthen the conclusions made by an examiner.

Due the systematic nature of the proposed ESS method, these comparative analyses are also more reproducible. An examiner, with no prior knowledge of a previous examination, can conduct a second examination of a comparison performed by a colleague by comparing each comparison area of an edge individually. By documenting alignment and any characteristic features found in the areas, the examiner could then calculate the ESS in relation to their examination, which should not be significantly different that the ESS reached by their colleague. If there is a disagreement, the documentation of the alignment of individual comparison areas would immediately highlight the areas along the edge where there is a difference in opinion. This permits a clear, constructive dialogue into the reasons behind the disagreement.

Throughout this study, the ESS method for both textiles and paper will be assessed by calculating performance rates, generating various graphical displays, and analyzing the dataset using Bayesian approaches, namely score likelihood ratios and evaluation of the evidence given two mutually exclusive propositions. Performance rates will be used to evaluate the overall rigorousness of the method, and will include such metrics as false positive rate, false negative rate, sensitivity, specificity, and overall accuracy. Theses metrics take into account the various outcomes of the comparative analysis process, of which there are five. These five outcomes include True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), and Inconclusive (IN). A true positive outcome occurs when an examiner concludes that a pair of samples "match" when the two samples are known to come from a common source. A true negative outcome is the result of a "non-match" conclusion when the samples are known to come from different sources. A false positive outcome occurs when the examiner incorrectly concludes that two samples "match" when, in reality, the samples come from different sources. A false negative outcome occurs when an examiner reaches a "non-match" conclusion when the samples are known to come from a common source. An inconclusive result can produce the correct expected answer, or be treated as a false positive or false negative result, depending on the particular design of the ground truth. As previously mentioned, performance rates can be calculated using the number of times these five outcomes occur in a dataset, and the equations for these performance rates can be viewed below.

Table 1. *Equations that will be used to calculate performance rates throughout this study. Equations for Accuracy, Sensitivity, Specificity, FPR and FNR are provided* [43].

Performance Rate (%)	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN + IN} \times 100$
Sensitivity	$\frac{TP}{TP + FN} \times 100$
Specificity	$\frac{TN}{TN + FP} \times 100$
False Positive Rate	$\frac{FP}{FP + TN} \times 100$
False Negative Rate	$\frac{FN}{FN + TP} \times 100$

Beyond performance rates, the dataset will be analyzed via box-plots, density functions, receiver-operator characteristic curves, and score likelihood ratios. Box-plots graphically display the spread of the data through the use of quartiles. These quartiles subdivide the data points based on five data points, including the median, maximum, minimum, first quartile, and third quartile. The median is the central point in the data set, while the maximum is the largest value and the minimum is the smallest value. The first quartile is calculated by finding the median for the values in between the median and the minimum and the third quartile is calculated by finding the median for the values in between the median and the maximum. A box is then drawn which encompasses the first quartile, the median, and the third quartile. Lines are then drawn extending out from the box to the minimum and maximum values, forming the "whiskers" of the box-and-whisker plot. These plots can be used to visualize the spread of the ESS metrics assigned to both the true positives and true negatives in the dataset. Similarly, density functions can also be used to visualize the distribution of scores for the dataset. These two methods are also useful for assessing overlap in the dataset between true positives and true negatives. This can occur when, for example, a score of 50 is given to both a true positive comparison and a true negative comparison.

Bayesian statistics will also be used for data analysis purposes [44]. This will come in the form of a score likelihood ratio that describes the likelihood of the observed evidence based on the ratio of the probabilities of each of two conflicting hypotheses being correct. Oftentimes in forensic science, these hypotheses come from the prosecution and defense. The prosecutorial hypothesis is therefore referred to as H_P, while the defense hypothesis is referred to as H_D. This results in the equation shown below.

$$Likelihood\ Ratio\ = \frac{p(H_P)}{p(H_D)}$$

The prosecutorial hypothesis is generally stated as “the defendant committed the crime” whereas the defense hypothesis is generally stated as “someone other than the defendant committed the crime”. In the case of this research, H_P and H_D will relate to the outcome of a comparison between two fractured pieces of a material. H_P will represent the hypothesis that the two pieces originated from the same source (a match), whereas H_D represents the hypothesis that the two pieces originated from different sources. This equation is shown below, where SLR stands for score likelihood ratio.

$$SLR\ = \frac{p(ESS\mid same\ source)}{p(ESS\mid different\ source)}$$

The probabilities used in the above equation are determined using the aforementioned density functions. Kernel Density Estimation can be used to generate a continuous function that can be used to estimate the probability of a specified score for both a true fit and a true non-fit. As the equation is a ratio, SLRs above 1 would indicate that there is more evidence to support a fit versus a non-fit. As the SLR increases far beyond 1, there is more and more evidence to support a “fit” conclusion. The opposite is true for SLRs less than 1. These would indicate the presence of evidence that supports a “non-fit” conclusion by an examiner. An SLR of exactly 1 would indicate equal evidence for a “fit” and a “non-fit”, which would likely result in an “inconclusive” decision. The base-10 logarithm of the SLR can also be calculated to make interpretation simpler. For log SLRs, positive values indicate evidence supporting a “fit” and negative values indicate evidence supporting a “non-fit”.

Receiver Operator Characteristic, or ROC, curves will also be used to evaluate the ESS metric for textiles and paper fracture comparisons. A ROC curve displays the false positive rate on the x-axis against the true positive rate on the y-axis. An ideal ROC curve is one that is “Γ”-shaped, indicating that the metric is effective at identifying, in this research, when two samples originated from the same source. This also indicates 100% specificity. On the other hand, a ROC curve can also be diagonally-shaped, which would indicate that the metric is useless at discriminating between a “match” and a “non-match”, and is, therefore, equally likely to conclude that a given comparison results in a “match” or a “non-match”. The accuracy of the method can also be determined using a ROC curve. This is done by calculating the area underneath the curve, or AUC. According to Fawcett, “the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance” [45]. It is important, however, that the dataset has somewhat equally-distributed matches and non-matches. For a dataset that is predominantly comprised of matches for instance, the metric’s accuracy may be lower than simply classifying every comparison as a “match”.

A logistic regression model was used to study the effects of certain factors on the quality of a physical fit for textiles. Logistic regression models are used to predict the probability of an event, expressed in log-odds, using different independent predictors. In this case, the predictors are factors that are thought to influence the quality of a textile physical fit, such as the separation method (hand-torn or stabbed), the construction of the textile (knit or weave), the composition of the textile (cotton, polyester, etc.), and the design of the textile (unicolor or multicolor). The model will determine if any interactions are present between the predictors. For example, a possible interaction could be one between construction and separation method, where knit hand-torn pairs result in substantially lower scores that other construction/separation method combinations. Ideally, this model would be consistent with previously-performed exploratory analysis.

In order to determine if interactions between predictors are present, model selection must be done. Multiple models are tested, with various predictors and interactions between predictors, by analyzing the number of parameters in the model, as well as the widely applicable information criterion (WAIC). When comparing models, the optimal model is the one with a relatively low WAIC value and low number of parameters. It is important to note that models can only be compared if they are modeling the same set of data. For example, if one model is using a dataset of 600 samples, whereas another model is using a dataset of 900 samples, then those models' WAIC values cannot be compared to each other.

Counterfactual plots are generated using the information from the model. These plots describe the change in output when the predictors are changed. In this case, the counterfactual plots will describe how the scores change when certain factors are changed, such as separation method, which will provide an understanding of how these different factors influence the edge similarity scores, and thus the quality of a physical fit for textiles.

3. DEVELOPMENT AND EVALUATION OF A SYSTEMATIC, QUANTITATIVE APPROACH FOR ASSESSING PHYSICAL FITS OF TEXTILES

3.1 OVERVIEW

Textiles are commonly encountered at crime scenes and can be relevant in assaults, murders, burglaries, and other cases. When a textile is fractured, the separated pieces can be realigned at the fractured edge to demonstrate that they were once joined together to form a single object. This is known as a physical fit and is considered the highest degree of association between two objects. Nonetheless, identifying a physical fit is subjective, and consensus best practices are still in development. This study introduces a systematic method for identifying and documenting relevant features in textile comparisons and assessing the quality of a physical fit using an edge similarity score. The proposed standardized terminology and documentation criteria in this method streamline the peer review process and simply convey the probative value of the evidence. A set of 967 blind comparisons were conducted to validate this method for textile physical fit analysis. The effect of the separation method (stabbed, hand-torn), design (unicolor, multicolor), and construction (knit, weave) did not show a substantial effect in overall edge similarity scores for fits and non-fits grounds-truth sets. However, fabric composition showed a considerable effect on the quality of a fit, with highly deformable polyesters being unsuitable for fit comparisons. Prior to physical fit examinations, a textile suitability assessment is recommended because highly deformable fabrics showed poor accuracy (61%). For those textiles deemed suitable for fit comparisons, the accuracy ranged from 88% to 100%, depending on the set. In total, ten false-positive misclassifications were reported from 477 possible true non-fitting comparisons (2%), indicating the importance of assessing the quality of a fit. The average ESS for a false positive comparison was 85%, and four of them received a score of 100%. On average, an ESS higher than 80% and an ESS lower than 40% supported a fit and non-fit decision, respectively. Experience and training were shown to improve examiner performance by approximately 7%. Overall, this study is anticipated to provide valuable tools for examining, interpreting, and reporting textile fracture fits, and providing a path forward.

3.2 INTRODUCTION

A physical fit, by definition, directly links two items as originating from the same object. This occurs when the characteristics and features of the material, such as a piece of clothing torn from a shirt, realign across the separated edges demonstrating they were once joined together. Because the items are directly linked, a physical fit is considered the highest degree of association between two objects. However, research is still needed to establish the reliability and validity of physical fit analysis[2,3]. This study addresses a recent call for research into the development of quantitative methods and the evaluation of error rates to assess physical fit comparisons by The National Institute of Justice (NIJ) Forensic Science Technology Working Group (TWG) and the

Organization of Scientific Area Committees (OSAC) recently published a call for research into the development of quantitative methods and the evaluation of error rates to assess physical fit comparisons [4,5], which this study aims to address.

Recent publications[8,46] and a survey from the NIST-OSAC Trace Materials Subcommittee Physical Fits Task Group[6] identify textiles, polymers, and paper as the items most frequently submitted to forensic laboratories for physical fit examinations. In the literature, several types of studies, including case studies, fractography research, and performance evaluations, demonstrate the value of physical fits for various materials. An early study focusing on glass fractography describes a qualitative comparison of glass chips to their original mediums through the analysis of hackle marks between the two samples[47]. In contrast, other studies highlight methodologies to compare matchbooks and fingernails[48,49].

Reports describing the use of fracture-matching in casework make up most of the literature surrounding physical fit analysis. A significant portion of this literature is devoted to metals, often taking the form of a toolmark examination. For instance, one case study describes the use of striations as a feature of interest for comparing the tip of a broken screwdriver found at the scene of a breaking-and-entering case with a broken screwdriver found in the suspect's vehicle. In another case[16], striations found in metal, in addition to edge morphology, were used to compare two halves of a broken knife used in a stabbing.

Other prominent materials that others have studied include plastics and paint. One report describes the features of interest that were useful in reaching a physical fit conclusion for soft plastic bags, including garbage bags. Edge striations were identified as a relevant feature, as were perforations and surface scratches[20]. Knowledge of the manufacturing process was critical in this study, as the orientation of specific markings on the garbage bags could be used to ascertain the production sequence[21]. Paint can be especially relevant in cases involving automobiles, and identifying a physical fit between a paint chip and a vehicle can link that vehicle to a crime. Case reports have highlighted the importance of layer and striation alignment when examining a potential physical fit for a paint chip[22]. In contrast, others discuss the impressions that can be left in the paint by welding beads, door frames, and other objects[23,24].

Quantitative and qualitative assessments of physical fits are also discussed in the literature, especially regarding duct and electrical tapes. One study by Bradley et al.[25] using duct tapes addresses examiner performance by describing accuracy rates for four examiners who conducted physical fit comparisons for several sets of duct tapes with known ground truth. The sets consisted of both hand-torn and scissor-cut tape pairs. While there were no false positive or false negative misclassifications, some pairs were classified as inconclusive. Additionally, a higher portion of scissor-cut pairs was classified as inconclusive (19%) compared to hand-torn pairs (8%). The same author conducted a separate study on electrical tape, using sets with variations in separation method, tape brand, and the person creating the tape pairs[26]. The ten sets contained 106 known true fits. Of these 106 true fits, 96 were correctly classified, while seven were deemed inconclusive, and one was misclassified as a false positive.

Tulleners and Braun laid the foundation of the present study in their 2011 study that outlined a method for assessing the quality of a duct tape physical fit[27]. The authors proposed a "percent match" metric that is calculated by measuring the distance of the matching area along the comparison edge in centimeters and dividing it by the total length of the comparison edge. A "match category" is also described that allows the examiner to express his/her confidence in a given physical fit by assigning it one of six classifiers. A follow-up study discussed examiner performance for a set of 1600 duct tape pairs, containing both torn and cut tapes[28]. The authors reported high accuracy rates (98% to 100%) and low misclassification rates (false-positive rates lower than 0.7 % and 3.3%, and false-negative rates lower than 2.7% and 0.3% for hand-torn and cut-tapes, respectively).

Prusinowski et al. built upon this work by defining a systematic approach to quantitatively assess duct tape physical fits[7]. In this study, the authors employed an Edge Similarity Score (ESS) to quantitatively define the quality of a physical fit between two pieces of duct tape. Duct tape is comprised of three general layers: the adhesive layer, the reinforcement layer, and the backing layer. The reinforcement layer, also known as scrim, is made of fibers oriented in both the warp and weft directions, forming a systematic way to subdivide the length of a duct tape fracture into comparison areas. Each scrim bin was defined as a comparison area. The ESS is a ratio between the number of matching comparison areas and the total number of bin comparison areas. The authors used this ESS metric to assess 2280 duct tape comparisons and reported high overall accuracy (84.9% to 99%) and no false positives. The study included various qualities of duct tape that were fractured using hand-torn and scissor-cut separation methods. It was found that ESS values higher than 80% were indicative of a match, while ESS values lower than 20% were indicative of a non-match.

To date, no such quantitative approaches exist for physical fit comparisons of textiles that formally assess performance rates for these materials or describe specific features and characteristics that could be useful in a physical fit comparison. However, the previously discussed research on duct tapes[7,27,28] provides a strong foundation for textiles due to some common characteristics between textiles and tape, such as the use of a woven reinforcement layer in duct tape that is of similar construction as a traditional textile. Despite this similarity, there are significant differences in the physical and chemical structures of textiles and duct tapes that require developing and evaluating a fit-for-purpose method.

Some understanding of features observed during separation can be derived from studies of damaged and severed fabrics. One fractography study discussed the type of textile damage observed after a stabbing event, detailing the differences in damage done by a dull blade compared to a sharpened blade and how fabric construction can impact the edge morphology of a fractured textile sample[30]. Another such study describes several more factors that can affect a fracture. In addition to the above elements, the authors also found that a serrated blade causes more distortion along the fracture edge than a straight blade. Interestingly, it was also found that fractures measured four hours after the stabbing event were found to be longer than when the same fractures were measured immediately after the event, implying that the yarns at the fractured edge begin to loosen and unravel as the time after the fracture event increases.

Despite efforts to develop protocols for physical fit examinations, they are still inherently subjective and thus susceptible to bias as the main instrument used in the physical fit comparisons is the human eye and brain. Thus, potential sources of bias in these examinations must be considered and limited. In recent studies, Dror et al. outline several sources of bias in forensic casework and various approaches to minimize them while increasing the repeatability, reproducibility, and transparency of forensic examiners' decisions[50], while Quigley-McBride et al. describe a method of limiting bias by using a linear sequential unmasking strategy for guiding examiners through their casework[51]. Several of these sources of bias are applied to the physical fit analysis and can be minimized by incorporating standardized protocols into the workflow.

This study addresses some of these issues by implementing a systematic method for comparative fit analysis that presents a more objective means of conducting textile comparisons. Here, standardized terminology, criteria, and descriptions for relevant features are established to facilitate reproducible data and transparent documentation and communication of results. This ensures that the relevant features are documented clearly across practitioners and facilitates blind peer review processes such as technical reviews and verifications. Also, the proposed method provides a means to assess the scientific validity of textile physical fit examinations and identify potential sources of error.

3.3 METHODS

3.3.1 Participant Training and Experience

The participants in this study were three graduate students within our research group. Prior to beginning the comparisons, each participant conducted a thorough review of relevant physical fit literature, which included physical fit studies of various materials and textile damage studies, and each had previously been trained on the ESS method for duct tape, including the identification of relevant features. The participants were required to complete an initial set of 40 duct tape comparisons, undergo a review session, and then complete an additional set of 80 duct tape comparisons. The participants' ESS results were compared to previously established consensus scores for each comparison pair to ensure accuracy. Participants A and B examined the suitability set, while Participants A and C examined the inter-examiner variability set. Participant C examined all subsequent sets. The participants in this study will be referred to as "examiners" for the remainder of this chapter.

3.3.2 Sample preparation

The textile dataset used in this study consisted of 967 textile fit comparisons from 774 paired items of various compositions, construction, design, and separation methods. A sampling diagram that breaks down the samples used to answer each question of interest can be seen in Figure 3.

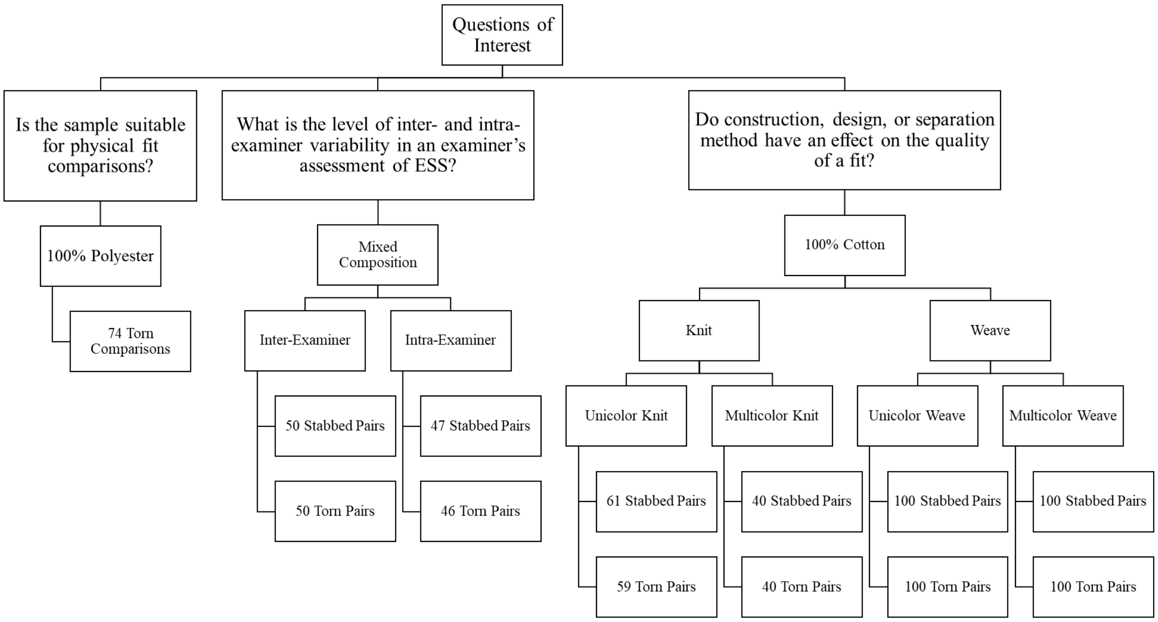


Figure 3. *Diagram of the distribution of the textile comparison sets by composition, construction, design, and separation method. The intra-examiner set used textile samples from the same set as the inter-examiner study.*

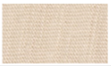
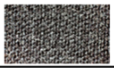
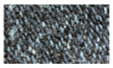
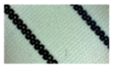
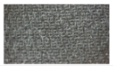
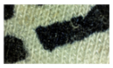
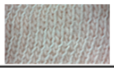
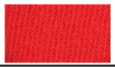
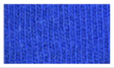
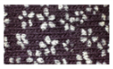
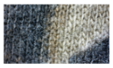
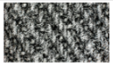
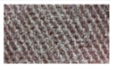
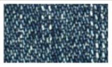
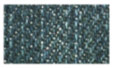
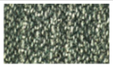
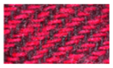
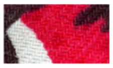
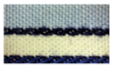
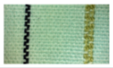
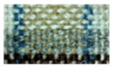
A preliminary suitability study was prepared to assess the application of the ESS method to textiles. Jersey knit, 100% polyester bolt fabric was used for this purpose. The fabric was cut into 100 rectangles, measuring approximately 26 cm in length and 18 cm in width. The fabric was separated in the width direction by hand-tearing aided by a central, 3 cm cut in the fabric. The samples were labeled and stored as all true-fitting pairs. They were then rearranged and relabeled into known true fitting and true non-fitting pairs by a separate person to keep the examiners blind to the ground truth of the set. The samples were ironed before analysis because curling was observed along the fractured edges. Two examiners independently conducted physical fit comparisons of the samples by subdividing the comparison edge into ten bins of equal size and performing individual, bin-by-bin comparisons. Of the 100 designed comparisons, only 37 comparisons were completed by each examiner due to substantial edge distortion.

Following the analysis of the preliminary study, a larger clothing dataset was prepared using different articles of clothing, incorporating fabrics such as cotton, rayon, and polyester. This study aimed to assess inter and intra-examiner variability in the ESS estimation. To that end, a new set of 100 comparison pairs was made, including 20 pairs from each of five garments. The clothing was placed on a mannequin carved out of foam (Foam Factory Inc.), approximately six inches thick, to represent a mock crime scenario. Two separation methods were employed: hand-tearing and stabbing. Ten hand-torn and ten stabbed comparison pairs were sampled from each article of clothing. The stabbing was performed at the height of 18 inches from the surface of the clothing using motion in the elbow only for the action to be as reproducible as possible. The opposite edges of the samples that were irrelevant for comparison were cut in a distinctive way to guide the examiner to analyze only the intended comparison edge. The pairs were scrambled and relabeled using a random number generator by a person not conducting the examination to keep the examiners comparing the samples blind to the ground truth.

To study intra-examiner variation, the set was relabeled with an entirely different numbering scheme and presented to the second examiner of the inter-examiner variation set under the pretense of being an entirely new set approximately three months after completion of the original set. Seven pairs determined to be too distinctive or memorable were removed from the set completely, resulting in a total of 93 comparison pairs. The examiner was not informed of the duplication or true identity of the set until after their examination was complete.

A third experiment was designed to determine the effect of different features on the quality of a physical fit. The features that were considered in this study include construction (knit or weave), design (unicolor or multicolor), and separation method (stabbed or torn). The composition of the fabric was kept fixed between all samples (100% cotton). The 600 cotton pairs were divided between the unicolor weave, multicolored weave, unicolor knit, and multicolor knit constructions. In this case, "unicolor" refers to a single-color fabric, such as an all-blue t-shirt. "Multicolor" refers to fabric with an all-over design or pattern, such as a camouflage shirt. Table 2 describes the fabrics used for each study, including their composition, construction, and image.

Table 2. *Table of fabrics used in this study, including their composition and construction, separated by set. The number in parenthesis in the description column represents the respective textile ID number.*

	Description	Composition	Construction	Image
Suitability	(1) Tan bolt fabric	100% polyester	Knit	
Inter- and Intra- Examiner Variability	(2) Navy dress pants	75% polyester, 25% cotton	Weave	
	(3) Navy denim jeans	60% cotton, 22% rayon, 17% polyester, 1% spandex	Weave	
	(4) White short-sleeve dress shirt with blue stripes	100% cotton	Weave	
	(5) Beige tank-top	100% polyester	Weave	
	(6) Navy and white patterned short-sleeve shirt	93% rayon, 7% flax	Knit	
Unicolor Knit	(7) Pink T-shirt	100% cotton	Knit	
	(8) Red T-shirt	100% cotton	Knit	
	(9) Blue T-shirt	100% cotton	Knit	
Multicolor Knit	(10) Navy polo shirt with white pattern	100% cotton	Knit	
	(11) Camouflage T-shirt	100% cotton	Knit	
Unicolor Weave	(12) Grey denim jeans	100% cotton	Weave	
	(13) Navy pants	100% cotton	Weave	
	(14) Light blue denim jeans	100% cotton	Weave	
	(15) Dark denim jeans	100% cotton	Weave	
	(16) Black Denim Jeans	100% cotton	Weave	
Multicolor Weave	(17) Red flannel lounge pants	100% cotton	Weave	
	(18) Pink penguin-patterned lounge pants	100% cotton	Weave	
	(19) Navy dress shirt with stripes	100% cotton	Weave	
	(20) Teal dress shirt with stripes	100% cotton	Weave	
	(21) Tan and black flannel shirt	100% cotton	Weave	

3.3.3 Comparison methods

Every examiner in this study used a EZ4 stereomicroscope (Leica Microsystems; Deerfield, IL). Examiners were free to use any level of magnification achievable using the stereomicroscope, though 35X magnification was found to be optimal for this study. The examiners independently compared each textile set using the ESS method by subdividing the edges of the comparison edge into ten bins of equal length, as illustrated in Figure 4. The ESS method quantifies the quality of a fit or non-fit and documents the examined areas of the sample, which allows for a straightforward peer review process, as two examiners can directly compare their results and notes for the same area of a given comparison. The examiners examined each bin along the comparison edge using a stereomicroscope and noted any distinctive features that strengthened or weakened their confidence in a particular decision. These features were defined as construction alignment, design alignment, edge alignment, yarn alignment, extreme distortion, secondary tearing, and fluorescence. Features present in each bin were documented, and the examiners noted when a specific feature was influential in determining a fit or a non-fit conclusion. An example of the documentation template is provided in appendix I.

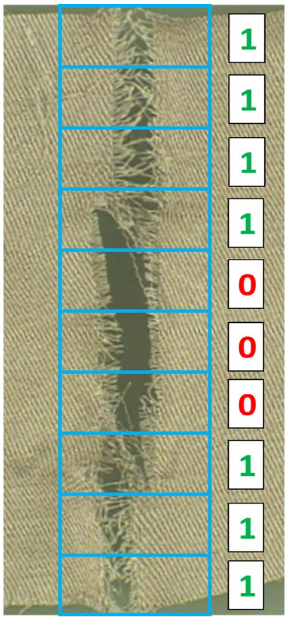


Figure 4. *An example of the systematic method of comparing two fractured textile samples. The examiner has identified four fitting bins, followed by three non-fitting bins, and finally three fitting bins, for an overall edge similarity score of 7/10 or 70%. This informs the examiner, who reached a low confidence fit decision.*

Each bin comparison received its own separate fit/non-fit classifier, either a 1, 0.5, or 0, for a fit, inconclusive, or a non-fit, respectively. This was then used to calculate the overall ESS for the compared edges. The examiner then assigned one of five qualitative classifiers to the comparison: F+, F-, INC, NF-, or NF+. The +/- attached to the fit and non-fit qualifiers allow the examiner to express their confidence level in the conclusion, or the quality of a fit/non-fit. For example, an F+ qualifier would indicate a fit with a high level of confidence and no documented limitations. This typically corresponded to an ESS of 80-100. An F- qualifier would indicate a fit with limitations and a lower degree of confidence, usually corresponding to an ESS of 60-80.

3.3.4 Data analysis

Performance rates were used to evaluate the method. False-positive rates, false-negative rates, sensitivity, specificity, and accuracy were calculated for each textile dataset. These metrics consider the five outcomes of the comparative analysis process, of which there are five. These five outcomes include true-positive (TP), true-negative (TN), false-positive (FP), false-negative (FN), or inconclusive (IN) results. A true-positive outcome occurs when an examiner concludes that a given pair of samples "fit" when the two samples are known to be a true fit. A true-negative outcome results from a "non-fit" conclusion when the samples are known to originate from different sources. A false-positive outcome occurs when the examiner incorrectly concludes that two samples "fit" when, in reality, the samples were not once from a single item. A false-negative outcome occurs when an examiner reaches a "non-fit" conclusion when, in fact, the samples were a true fit. An inconclusive result can produce the correct expected answer, or it can be treated as a false-positive or a false-negative result, depending on the particular design of the ground truth. Performance rates are calculated using the number of times these five outcomes occur in a dataset, and the equations used for these rates can be seen in Table 1 (page 26).

Beyond performance rates, the data were also analyzed using box plots and logistic regression models. Box plots graphically display the spread of the data through quartiles, which subdivide the data based on five data points: the median, maximum, minimum, first quartile, and third quartile. These plots can aid in visualizing the spread of the ESS metrics assigned to true-positives and true-negatives in the dataset, as well as any potential overlaps between ground truth sets. The dataset can also be subdivided into classes based on different characteristics, such as construction type. Boxplots of these classes can inform the given characteristic's possible effects on the scores. They help assess overlap in the dataset between true-positives and true-negatives. This can occur when, for example, a score of 50 is given to both a true-positive comparison and a true-negative comparison.

A logistic regression model was used to study the effects of certain factors on the ESS score and thus, the quality of a physical fit for textiles. Logistic regression models are used to predict the probability of an event, expressed in log-odds, using different independent predictors[52]. In this case, the predictors are factors that are thought to influence the quality of a textile physical fit, such as the separation method (hand-torn or stabbed), the construction of the textile (knit or weave), the composition of the textile (cotton, polyester, etc.), and the design of the textile (unicolor or multicolor). The model will determine if any interactions are present between the predictors. For example, a possible interaction could be one between the construction and separation method, where knit hand-torn pairs may result in substantially lower scores than other construction/separation method combinations. Ideally, this model would help to explain the observations from the exploratory analysis.

To determine if interactions between predictors are present, model selection must be made. Multiple models are tested, with various predictors and interactions between predictors, by analyzing the number of parameters in the model and the widely applicable information criterion

(WAIC). When comparing models, the optimal model is the one with a relatively low WAIC value and a low number of parameters. It is important to note that models can only be compared if they model the same data set. For example, suppose one model is using a dataset of 600 samples, whereas another model is using a dataset of 900 samples. In that case, those models' WAIC values cannot be compared to each other.

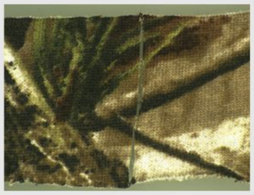
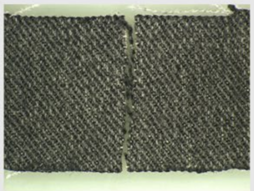
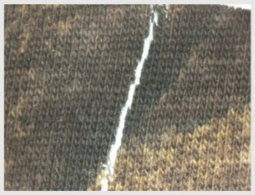
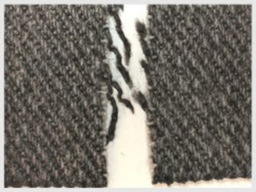
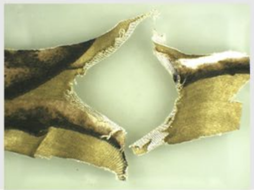
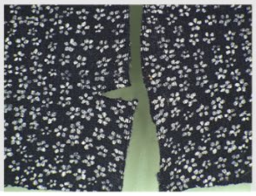
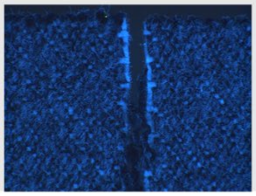
Counterfactual plots are generated using the information from the model. These plots describe the change in output when the predictors are changed. In this case, the counterfactual plots will describe how the scores change when certain factors are changed, such as the separation method, which will provide an understanding of how these different factors may influence the edge similarity scores and thus the quality of a physical fit for textiles.

3.4 RESULTS AND DISCUSSION

3.4.1 Standardization of relevant features

To create a systematic method for physical fit comparisons of textiles, it was essential to identify the textile-relevant features that can influence an examiner's decision-making process. Additionally, standardization of terminology and descriptors of these features makes reviewing and discussing comparisons between examiners more straightforward and transparent, as all examiners should recognize the same features and use the same terminology to reference these features. To that end, the examiners in this study documented the specific features and characteristics of the textile samples that influenced their opinion regarding a potential physical fit or non-fit. In total, seven prominent features were identified and defined, which are shown in Table 3.

Table 3. *Prominent features and terminology that can be useful for the determination of a physical fit*

Design Alignment	Construction Alignment	Edge Alignment	Yarn Alignment
Consistency and alignment of yarn color and pattern between two textile fragments	Consistency and alignment of construction, including type (weave/knit) and direction, between two textile fragments	Overall edge shape alignment. Identified edge shapes include straight, wavy, and puzzle-like	Alignment of yarns that have been pulled out of the fracture edge between two textile fragments
			
Extreme Distortion	**Secondary Tearing**		**Fluorescence**
Force applied during the fracture event causes distortion that can mask other features	A secondary, perpendicular tear that is not the primary fracture that is being compared		Fluorescence of individual yarns can aid in the identification of a physical fit
			

Some of these features are intrinsic to the fabric itself. This includes design alignment, construction alignment, and fluorescence. Design alignment refers to the agreement and alignment of fabric design across the comparison edge of two samples. An example of this feature could be the alignment of stripes of the same color and size or the continuation of a repeating or distinctive pattern on the textile. Construction alignment occurs when there is an agreement between two samples regarding the type and direction of the construction of the fabric. Two true fitting samples woven in a diagonal direction relative to the comparison edge would possess this feature. Finally, the fluorescence of individual yarns in the fabric can also aid in identifying a physical fit. These features typically increase an examiner's confidence in the presence of a physical fit, as they indicate that the two samples could have originated from one common source. However, fluorescence was rarely the determining factor for fits or non-fits.

Other features are extrinsic to the fabric and caused by the separation event. These features include edge alignment, yarn alignment, extreme distortion, and secondary tearing. Edge alignment denotes the alignment of the overall edge shape between two samples. Three common edge shapes were identified throughout this study and were denoted as straight, wavy, and puzzle-like. A straight edge consists of a comparison edge that is entirely perpendicular or angled to the upper and lower edges of the sample. A wavy edge possesses curves and dips along the fracture edge, while a puzzle-like edge exhibits generally large protrusions, similar to a jigsaw puzzle piece, that correspond with a true-fitting mate. The overall edge shape must align between the two fragments for a physical fit to occur. The presence of two radically different edge shapes in a comparison may indicate a non-fit. Yarn alignment refers to the alignment of loose yarns that have been pulled out of the fractured edge of a sample. This is much more common in hand-torn samples subject to vigorous pulling and tearing, which is absent from stabbed samples. Hand-torn samples are also much more likely to exhibit extreme distortion. This feature can mask other present features through distortion and stretching of the comparison edge during the fracture event. This feature can hinder a comparison and result in a "not suitable for comparison" declaration in severe cases because the comparison edge is too deformed for an accurate comparison. Secondary tearing is another feature that can mislead to a non-fit, though to a much lesser extent. A secondary tear describes a minor fracture, often perpendicular to the comparison edge, that is not the primary fracture being compared between two samples. This feature may cause a "non-fit" decision for a given bin, as the fracture will most likely only be present on one edge.

Understandably, some features are more common than others. Construction alignment, for example, is applicable for all comparisons, and statements about design alignment can be made for all cases involving multicolor fabric. On the other hand, secondary tearing is rare, only occurring in about 1% of comparisons in this study. By standardizing these features, consistent reporting between examiners can be established, improving the peer review process and minimizing the variation of terms used or documentation between examiners. This is a crucial first step in creating a systematic and more objective method for physical fit comparisons.

3.4.2 Suitability assessment

Due to differences in fabric composition and construction, recognition that not all fabrics are suitable for physical fit comparisons is critical. Some fabrics may produce an unreasonably large number of misclassifications compared to other fabrics. This is illustrated in the examiners' performance of the initial preliminary textile set, which was comprised of hand-torn, jersey knit, polyester fabric. Each examiner completed only 37 of 100 intended comparisons (74 overall) before a third person who knew the ground truth evaluated their independent assessments. It was determined that the overall accuracy for both examiners was too poor to continue with the current approach, as seen in Table 4. While neither examiner reached a false-positive conclusion, their combined false-negative rate was deemed unacceptable (63%). More than half of the total true-fitting pairs were misclassified as non-fits by the two examiners. Despite attempts to iron the fabric before comparison, curling at the comparison edge and overall distortion of the polyester knit fabric resulted in significant disagreement between the two examiners and high error rates. Regarding the overall conclusion, in terms of fit or non-fit, the examiners disagreed on 30% of the comparisons.

Table 4. *Performance rates for the preliminary textile set of highly deformable fabric*

Preliminary Textile Set	Reported Fit	Reported Non-Fit	Reported Inconclusive	Total Comparisons
True Fit	17 (37% True Positive)	29 (63% False Negative)	0 (0% Inconclusive)	46
True Non-Fit	0 (0% False Positive)	28 (100% True Negative)	0 (0% Inconclusive)	28
Accuracy	61%			

An example of a common false-negative misclassification can be seen in Figure 5. Due to stretching, curling, and yarn distortion at the fracture edge, the comparison was classified incorrectly as a non-fit. These findings raise awareness that fabric-type suitability criteria are necessary for physical fit examinations. Some fabrics under certain conditions may not be suitable for physical fit comparisons, as they are prone to extreme distortion and produce a substantial number of misclassifications, a significant number of which are not replicable. Among the remaining datasets investigated in this study, 100% polyester knit was the only fabric type that led to suitability issues. Thus, it is recommended to first assess the fabric distortion level. If the items are deemed unsuitable for a physical fit examination, the textiles must instead be considered for other chemical and physical comparisons.

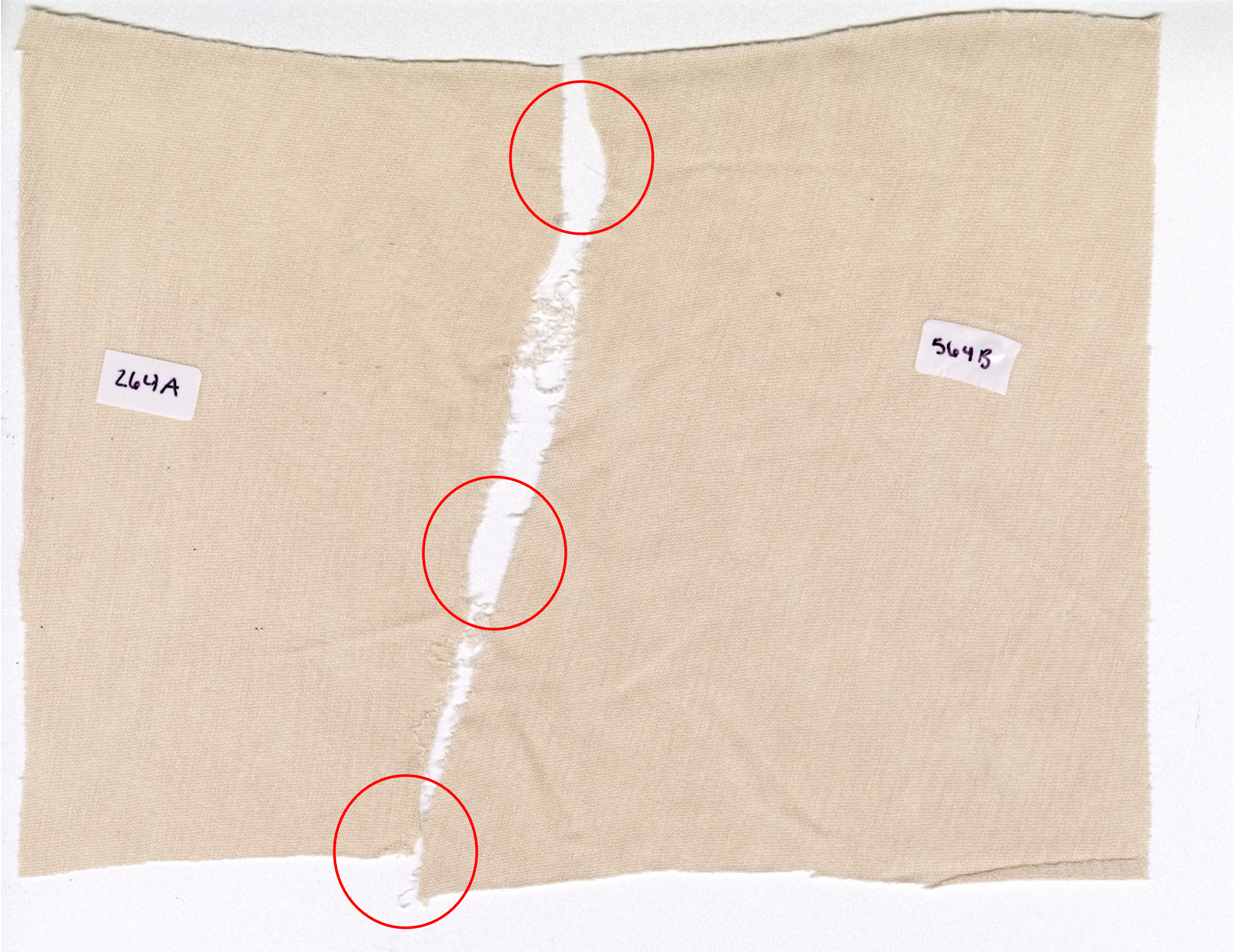


Figure 5. *False-negative misclassification from the preliminary textile set (Fabric 1, 100% polyester). The examiners noted the misaligned edge and inconsistent edge shape as the reasons for classifying this comparison as a non-fit. These areas are circled in red. A difference in the length of the edges is also observed due to stretching and distortion.*

3.4.3 Examiner variation

The inter-examiner variability was studied by creating a textile sample set using fabrics of various compositions, constructions, and designs which was examined by two examiners independently. Results for this study were promising, indicating low inter-examiner variability when using the edge similarity score estimation and reporting template to document the findings. As seen in Table 5, the accuracy rate was comparable between the two examiners for the stabbed subset. Both examiners misclassified four total comparisons, though Examiner 2 classified two additional comparisons as inconclusive, contributing to the lower accuracy. Two of Examiner 1's false-positive comparisons were reproduced by Examiner 2, while Examiner 1's third false-positive was classified as inconclusive by Examiner 2. All three comparisons that produced false-positive classifications were 100% polyester woven fabric, which reinforces the need to consider fabric suitability when evaluating the possibility of a physical fit between two items.

Table 5. *Performance rates for the inter-examiner variability textile set*

Inter-Examiner Variability Textile Set	Examiner 1 Reported Fit	Examiner 1 Reported Non-Fit	Examiner 1 Reported Inconclusive	Examiner 2 Reported Fit	Examiner 2 Reported Non-Fit	Examiner 2 Reported Inconclusive	Total Comparisons
	Stabbed Subset						
True Fit	25 (96% True Positive)	1 (4% False Negative)	0 (0% Inconclusive)	23 (88% True Positive)	2 (8% False Negative)	1 (4% Inconclusive)	26
True Non-Fit	3 (12% False Positive)	21 (88% True Negative)	0 (0% Inconclusive)	2 (8% False Positive)	21 (88% True Negative)	1 (4% Inconclusive)	24
Accuracy	92%			88%			
	Torn Subset						
True Fit	25 (96% True Positive)	1 (4% False Negative)	0 (0% Inconclusive)	21 (81% True Positive)	3 (11% False Negative)	2 (8% Inconclusive)	26
True Non-Fit	0 (0% False Positive)	24 (100% True Negative)	0 (0% Inconclusive)	1 (4% False Positive)	23 (96% True Negative)	0 (0% Inconclusive)	24
Accuracy	98%			88%			

An example of one of the false-positive comparisons that both examiners misclassified can be seen in Figure 6. The comparison areas determined to "fit" by both examiners are outlined in blue. Examiner 1 determined that two additional areas, Bins 1 and 10, also fit, and those areas are outlined in green. For a misclassification, the agreement between the two examiners is remarkable. It highlights one of the strengths of this method: transparency in the decision-making process and peer review. It is straightforward to identify the specific areas that each examiner considered during the comparison process by comparing their ESS documentation and bin-by-bin observations side-by-side. In this case, both examiners noted the importance of construction alignment when reaching their ultimately incorrect fit conclusion.

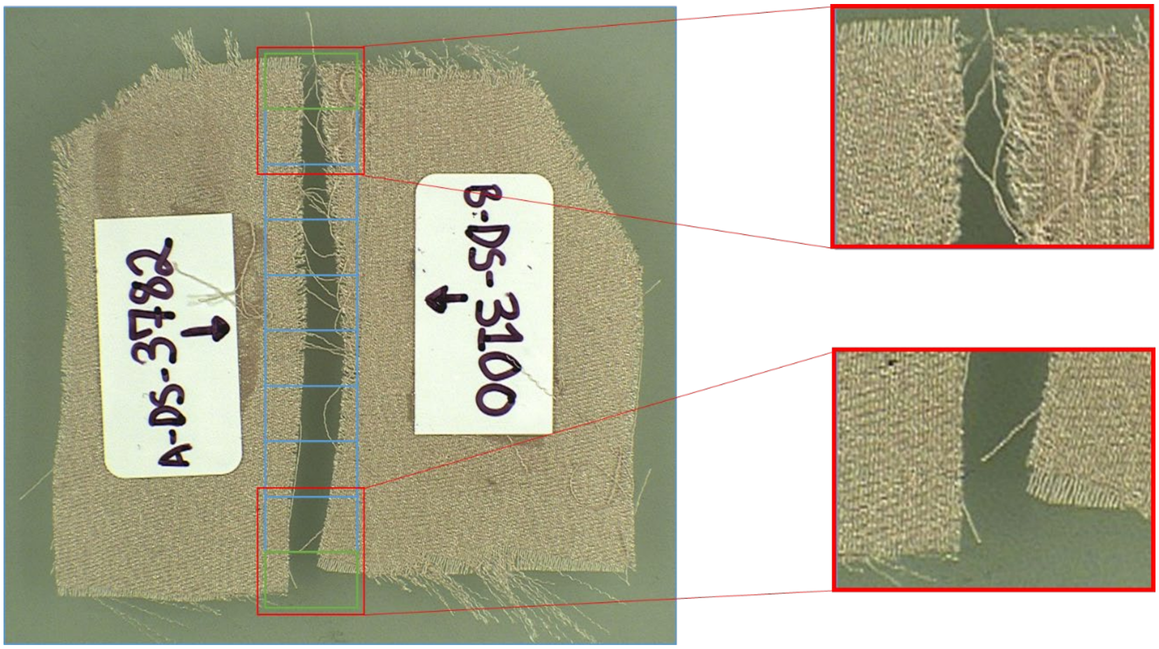


Figure 6. *An example of a true non-fitting comparison classified as a false-positive by both examiners (Fabric ID 5, 100% polyester). This pair was created from Fabric 5. Areas that were classified as a fit by both examiners are outlined in blue, while areas that were classified as a fit by Examiner 1 only are outlined in green. Some magnified areas of interest are showcased in red boxes on the right.*

The hand-torn subset also exhibited similar results regarding inter-examiner variability. Examiner 1's accuracy was 98% for this subset, while Examiner 2's accuracy was 88%. The hand-torn subset produced a smaller number of false-positive misclassifications than the stabbed subset, though there was a slight increase in the number of false-negative misclassifications. Examiners 1 and 2 agreed on the single misclassification made in the hand-torn subset. Both examiners assigned an ESS value of "0" to a true-fitting comparison. This incorrect non-fit classification occurred for a 100% polyester fabric that was fractured using a hand-torn separation method. When inspecting this pair, notable distortion was observed along the fracture edge, which was not seen in stabbed pairs.

Intra-examiner variability was studied by presenting Examiner 2 with the same comparison pairs, only relabeled, under the pretense of a new, completely unrelated set. This occurred approximately three months after the completion of the original set. The student was informed at a progress report meeting that a new dataset was created to increase the sample size to minimize possible suspicion of the dataset created for intra-examiner assessment. Interestingly, for both the stabbed and hand-torn subsets, examiner accuracy improved for the replicate dataset, shown in Table 6.

Table 6. *Performance rates for the intra-examiner variability textile set*

Intra-Examiner Variability Textile Set	Reported Fit Replicate 1	Reported Non-Fit Replicate 1	Reported Inconclusive Replicate 1	Reported Fit Replicate 2	Reported Non-Fit Replicate 2	Reported Inconclusive Replicate 2	Total Comparisons
	Stabbed Subset						
True Fit	20 (87% True Positive)	2 (9% False Negative)	1 (4% Inconclusive)	22 (96% True Positive)	1 (4% False Negative)	0 (0% Inconclusive)	23
True Non-Fit	2 (8% False Positive)	21 (88% True Negative)	1 (4% Inconclusive)	1 (4% False Positive)	22 (92% True Negative)	1 (4% Inconclusive)	24
Accuracy		87%			94%		
	Torn Subset						
True Fit	19 (82% True Positive)	2 (9% False Negative)	2 (9% Inconclusive)	23 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	23
True Non-Fit	1 (4% False Positive)	22 (96% True Negative)	0 (0% Inconclusive)	2 (9% False Positive)	21 (91% True Negative)	0 (0% Inconclusive)	23
Accuracy		89%			96%		

When looking at individual sample comparisons, it was noted that, generally, true matches received higher scores and true non-matches received lower scores for the second time this set was analyzed. This is likely due to increased experimental skills after the analyst became more familiarized with physical fit examinations, highlighting the impact of training and experience when conducting these examinations. This data shows that experience can increase an examiner's ability to conduct physical fit comparisons. One exception to this can be seen in Figure 7, which shows a true non-fitting comparison that was classified correctly as non-fit for replicate 1 but was misclassified as a fit for replicate 2. For replicate 2, this comparison was assigned an edge similarity score of 70 by the examiner, who noted construction alignment as a particularly influential feature for this comparison, leading to a false positive. This brings an important example of using the ESS to inform the examiner's opinion. The data shows that an ESS score below 80 does not provide strong support for a fit; therefore, an F- was reported by the examiner. In a case, this can be clearly expressed and prevent misleading evidence when using the ESS criteria. Given a physical fit's probative value, we recommend reporting a fit only for scores 80 or above for textiles. For lower scores, and in the absence of exclusionary differences, we recommend reporting a non-fit and submitting the items for chemical and physical textile/fiber comparisons, if appropriate. Here, for purposes of the performance rates evaluation, we used the threshold of 0-40 (non-fit), 40-60 (inconclusive), and 60-100 (fit) to encompass worst-case scenarios. Further discussion of these thresholds will be discussed in more detail in a later section.

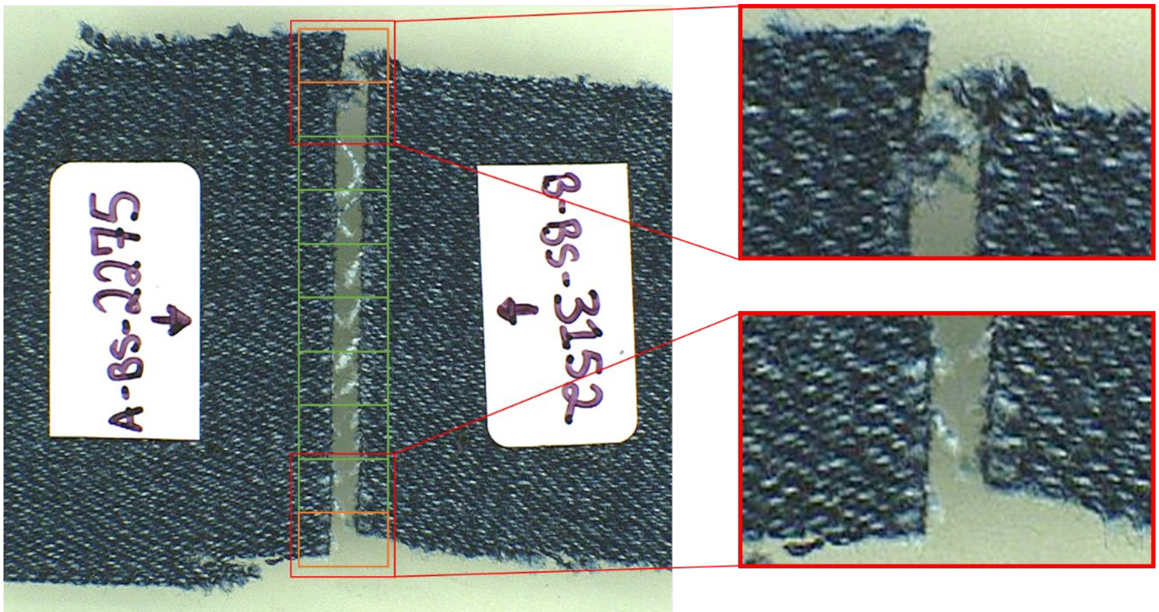


Figure 7. *Example of a true non-fitting comparison classified as a "fit" by the examiner for replicate 2 (Fabric ID 3, denim). Areas outlined in green were considered a fit for replicate 2 only, while areas outlined in orange were classified as a non-fit for both replicates. Areas of interest are showcased in red magnification boxes.*

3.4.4 Effect of factors on the quality of a physical fit

A final study was performed to quantify the effect of certain factors on the quality of a physical fit. Because composition had already been evaluated through the previous sets, it was held constant for all remaining comparison pairs. The 100% cotton was selected as the set composition due to its prevalence in everyday life and its tendency not to deform upon fracturing substantially. The factors considered in this study include construction (weave/knit), design (unicolor/multicolor), and separation method (stabbed/hand-torn).

The unicolor knit textile set (N=120) results can be seen in Table 7. The examiner correctly classified all comparison pairs in this set except for one true-fitting pair in the hand-torn subset that was deemed inconclusive. This hand-torn pair exhibited extreme distortion in some areas that hindered the comparison process but also displayed some areas of construction alignment, which resulted in the inconclusive decision. In comparison, results from the multicolor knit textile set (N=80) can be seen in Table 8. The accuracy rates for each set are strikingly similar. In both cases, accuracy for the stabbed subsets was 100%, while accuracy for the torn subsets was 98%. One true-fitting pair was classified as a non-fit, causing a false-negative result for the multicolor knit set. In the examiner's bin-by-bin records, they commented that the edges of the samples did not align, which caused a gap at the center of the fracture. This gap prevented the central bins from fitting together, leading to a non-fit decision. As is, these results do not indicate a substantial difference in performance between unicolor and multicolor knit fabrics, nor a substantial difference in performance for stabbed and hand-torn separation methods for 100% cotton textiles.

Table 7. *Performance rates for the unicolor knit textile set*

Unicolor Knit Textile Set	Reported Fit	Reported Non-Fit	Reported Inconclusive	Total Comparisons
	Stabbed Subset			
True Fit	30 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	30
True Non-Fit	0 (0% False Positive)	31 (100% True Negative)	0 (0% Inconclusive)	31
Accuracy	100%			
	Torn Subset			
True Fit	29 (97% True Positive)	0 (0% False Negative)	1 (3% Inconclusive)	30
True Non-Fit	0 (0% False Positive)	29 (100% True Negative)	0 (0% Inconclusive)	29
Accuracy	98%			

Table 8. *Performance rates for the multicolor knit textile set*

Multicolor Knit Textile Set	Reported Fit	Reported Non-Fit	Reported Inconclusive	Total Comparisons
	Stabbed Subset			
True Fit	20 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	20
True Non-Fit	0 (0% False Positive)	20 (100% True Negative)	0 (0% Inconclusive)	20
Accuracy	100%			
	Torn Subset			
True Fit	19 (95% True Positive)	1 (5% False Negative)	0 (0% Inconclusive)	20
True Non-Fit	0 (0% False Positive)	20 (100% True Negative)	0 (0% Inconclusive)	20
Accuracy	98%			

Woven 100% cotton textiles were also evaluated. Again, performance rates were assessed for fabric design and separation method. The unicolor weave textile set (N=200) results can be seen in Table 9. In this case, one false-negative was reported for the stabbed subset, resulting in an accuracy of 99% for this subset. The examiner reported an inconclusive result and a false-negative result for the torn subset, resulting in an accuracy of 98% for that subset. It was expected that the performance slightly worsens for the torn subset compared to the stabbed subset due to the higher amount of distortion and curling typically observed in hand-torn samples. Table 10 displays the results for the multicolor weave textile set (N=200). Interestingly, while the examiner achieved 100% accuracy for the torn subset, the examiner classified one true non-fitting comparison as a fit, resulting in a false-positive misclassification for the stabbed subset. An image of this comparison pair can be seen in Figure 8. This figure shows comparison areas the examiner deemed a fit in green and non-fit in red. Highlighted in blue are two regions of interest for this comparison. This example is notable in how well the penguin design on the fabric aligns. When combined with the construction and edge alignment seen in the upper zoomed-in window, this illustrates a situation that requires examiners to be cautious, as these features may cause the examiner to overlook the slight difference in the pattern that would indicate the presence of a non-fit. The lower zoomed-in window highlights an area of slight dissimilarity, as the white yarns comprising the penguin's lower torso do not quite align between the two samples. While both samples shared the same general edge shape, wavy, the edges did not fit together perfectly. The lower inset window highlights a slight discrepancy in the wave pattern, which could also be used to differentiate these samples and render a non-fit decision. However, these bins also presented some similarities that made this comparison pair particularly challenging.

Table 9. *Performance rates for the unicolor weave textile set*

Unicolor Weave Textile Set	Reported Fit	Reported Non-Fit	Reported Inconclusive	Total Comparisons
	Stabbed Subset			
True Fit	50 (98% True Positive)	1 (2% False Negative)	0 (0% Inconclusive)	51
True Non-Fit	0 (0% False Positive)	49 (0% True Negative)	0 (0% Inconclusive)	49
Accuracy	99%			
	Torn Subset			
True Fit	48 (98% True Positive)	1 (2% False Negative)	0 (0% Inconclusive)	49
True Non-Fit	0 (0% False Positive)	50 (98% True Negative)	1 (2% Inconclusive)	51
Accuracy	98%			

Table 10. *Performance rates for the multicolor weave textile set*

Multicolor Weave Textile Set	Reported Fit	Reported Non-Fit	Reported Inconclusive	Total Comparisons
	Stabbed Subset			
True Fit	47 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	47
True Non-Fit	1 (2% False Positive)	52 (98% True Negative)	0 (0% Inconclusive)	53
Accuracy	99%			
	Torn Subset			
True Fit	48 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	48
True Non-Fit	0 (0% False Positive)	52 (100% True Negative)	0 (0% Inconclusive)	52
Accuracy	100%			

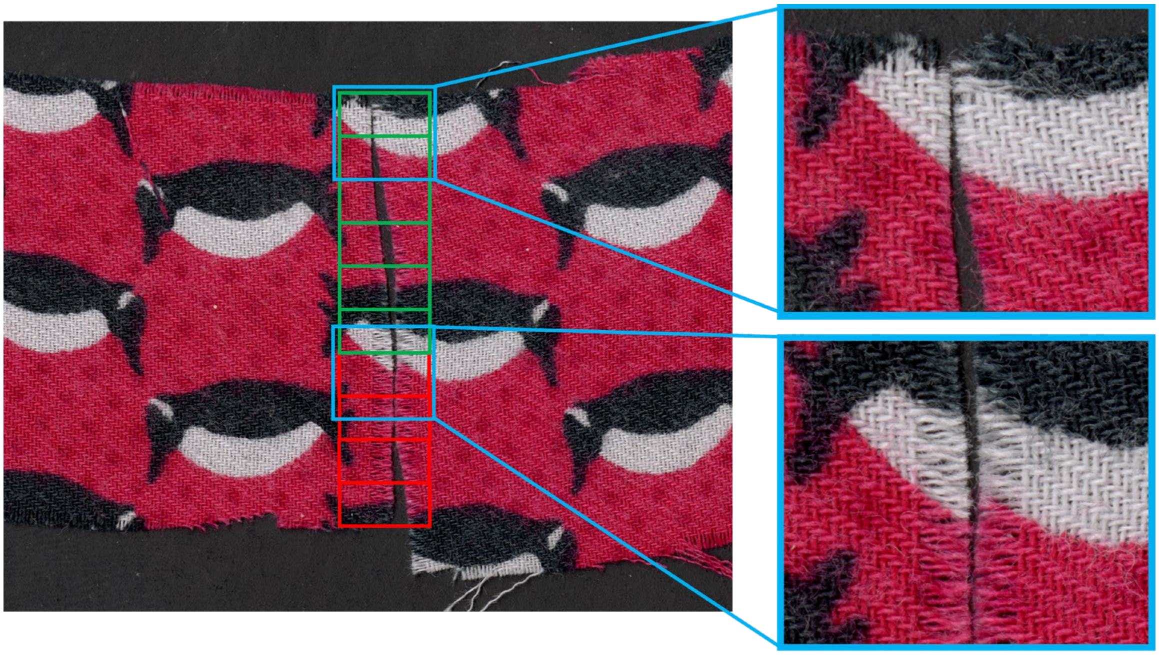


Figure 8. *False-positive comparison of a non-fit pair identified as a fit (Fabric ID 18, 100% cotton). This pair received an edge similarity score between 60 and 70% (weak positive). The bins identified as a fit by the examiner are outlined in green, while bins identified as a non-fit are outlined in red. The blue magnification boxes highlight two interesting areas of multicolored design alignment, one of the critical features cited by the examiner as influential in their decision. The examiner also noted the consistency in the weave direction as another feature of influence. There is a slight difference in edge shape in this comparison, however. While both samples are wavy in edge shape, there is a difference in the wave pattern.*

When considering each of the subsets and potential effects on the ESS and the quality of a physical fit, it is difficult to make any inferences using performance rates solely, as the rates are similar between subsets. However, differences appear when the edge similarity scores are analyzed more closely. Figure 9 shows the spread of the scores for true fits and true non-fits for each subset as boxplots. Overall, true fits appear to produce a broader range of scores (70-100) than true non-fits, which cluster at lower scores (0-10). This indicates that the examiner was more comfortable classifying a comparison as a non-fit. In contrast, for true fits, certain features influenced the score assigned to the physical fit identified by the examiner, producing a wider range of scores. When considering specific features that could affect the ESS, the separation method is a prominent one. It was hypothesized that hand-torn comparison pairs would be assigned a wider range of scores than stabbed comparison pairs because of the intense pulling and stretching during the tearing process, which is less likely to occur during the stabbing process. This stretching can distort the comparison edge of a sample, which can affect the features observed when conducting the comparison. Distortion can mask key features or potential areas of alignment, diminishing the quality of a fit and lowering the scores assigned to a true fitting comparison pair. Considering the spread of the scores, as hypothesized, hand-torn pairs received a wider range of scores than stabbed pairs from the same subset. However, stabbed comparisons can produce more false-positive classifications than torn comparisons, as the stabbing mechanism produces less distinctive edge patterns than the tearing process, which may make the identification of a non-fit slightly more difficult in some cases.

Fabric construction is another element hypothesized to affect the ESS for a physical fit. When considering the results of our previous studies, it was hypothesized that knit fabric would generally produce lower scores than woven fabric. This is because the knit fabric is more likely to unravel, stretch, and deform when fractured than woven fabric, which seems to hold its form better. This was supported by the scores, especially for the torn subsets. For true fits, hand-torn knit fabric produced the widest range of scores. In fact, the average score for the hand-torn multicolor knit fabric was approximately 70, while the average score for the hand-torn unicolor knit fabric was about 90. For all other subsets, the average score for a true fit was closer to 100.

On the other hand, the woven fabric may be prone to producing false positives at a higher rate than knit fabric because of the same reason. The tendency of woven fabric to deform less than knit fabric, especially in the case of stabbed samples, means that the comparison edges can be less distinctive. To reach their conclusions, examiners must consider other features in these cases, especially design and construction direction alignment.

Finally, the presence of designs on the fabric was also thought to influence the quality of a physical fit. The multicolor fabric was expected to increase the quality of a physical fit because the alignment of a design, such as a camouflage pattern, can cement an examiner’s conclusion that a physical fit is present. Unicolor fabric was expected to produce more misclassifications because of the absence of a design that could be compared and potentially aligned between two fragments. However, this was not observed experimentally, as there were no appreciable differences in the spread of the scores between unicolor and multicolor fabrics. Unicolor fabric produced more misclassifications, which is expected because these fabrics lack the multicolor design elements that can make identifying fits or non-fits more straightforward.

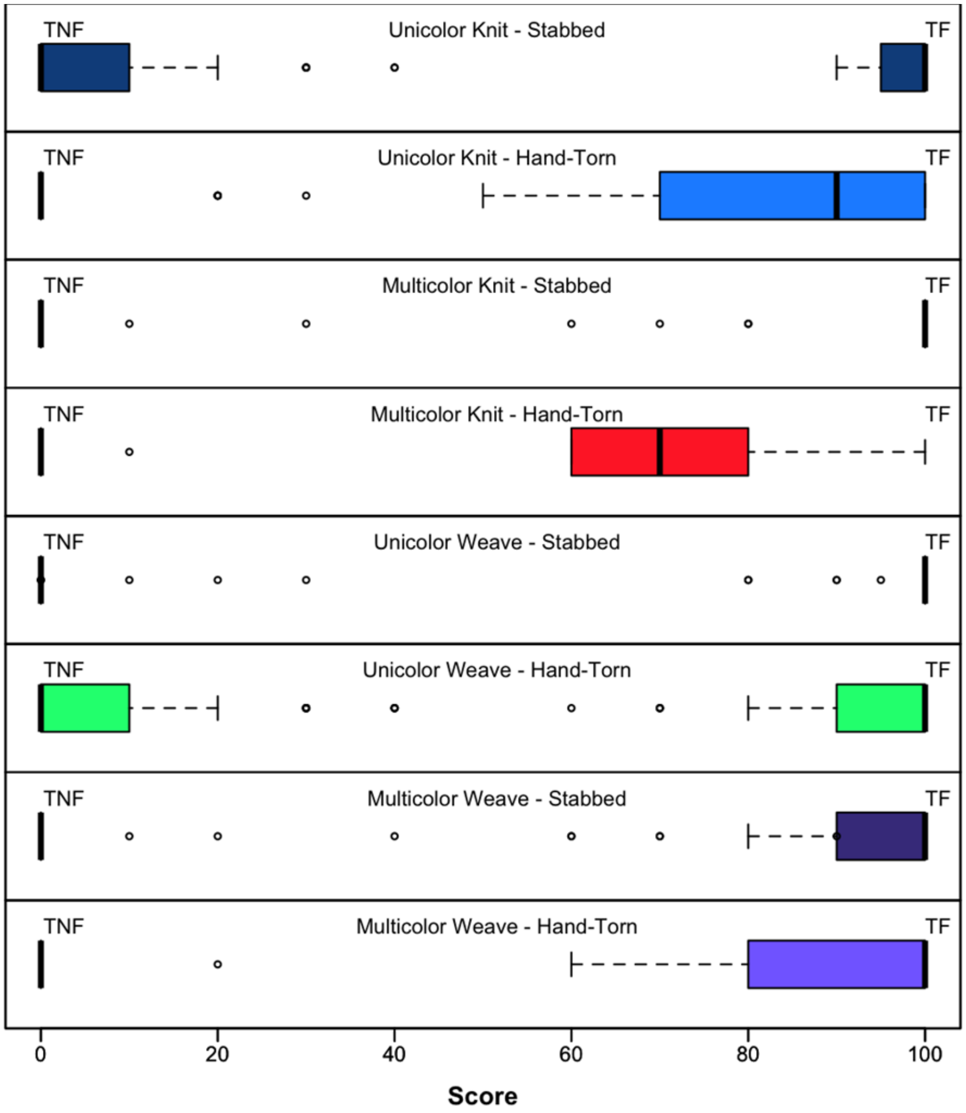


Figure 9. *Boxplots showing the distribution of edge similarity scores for each subset of textile comparisons. Scores for true non-fits are shown on the left, and scores for true fits are shown on the right.*

Overall, a good separation of ESS scores was observed between true fit and true non-fit comparison sets. However, the limitation of boxplot evaluations is that it does not allow for the assessment of any potential interactions between parameters. To better evaluate the influence of these factors on a physical fit, a logistic regression model was used to show the effect of each factor on the resulting edge similarity score. The equation used for the model incorporated the separation method, construction, and design. After evaluating several possible models, one was selected that includes an interaction between construction and separation method, which can be seen below. Model selection was done by evaluating the widely applicable information criterion (WAIC), as well as the number of parameters. WAIC provides an estimate of model quality by representing the loss of information by a model when processing data; it also accounts for both overfitting and underfitting. The simplest model with the lowest number of parameters and low WAIC value was selected. A potential interaction between separation method and construction was observed in the experimental data, as hand-torn knit comparison pairs seemed to produce a broader range of scores for true-fitting comparisons than other combinations of separation method and construction. This indicates that these two factors may have an additional interaction and influence on the ESS.

$$\log odds \sim -1 + Ground\ Truth * (Print + Construction * Separation\ Method)$$

From the logistic regression model, counterfactual plots were made to compare the effect of each factor of interest on the scores and allowed for exploration of the potential interactions between experimental factors that may not be readily apparent when observing boxplots or other exploratory metrics. In addition, parameter plots were generated to evaluate the magnitude of each factor. The counterfactual plots shown in Figures 10 and 11 demonstrate the effect of the construction and separation method, respectively, on the edge similarity scores. On the left side of the counterfactual plot, the pseudo data demonstrates simulated data generated by the model. The experimental data on the right represents the physical fit comparison ESS results reported by the examiners in this study.

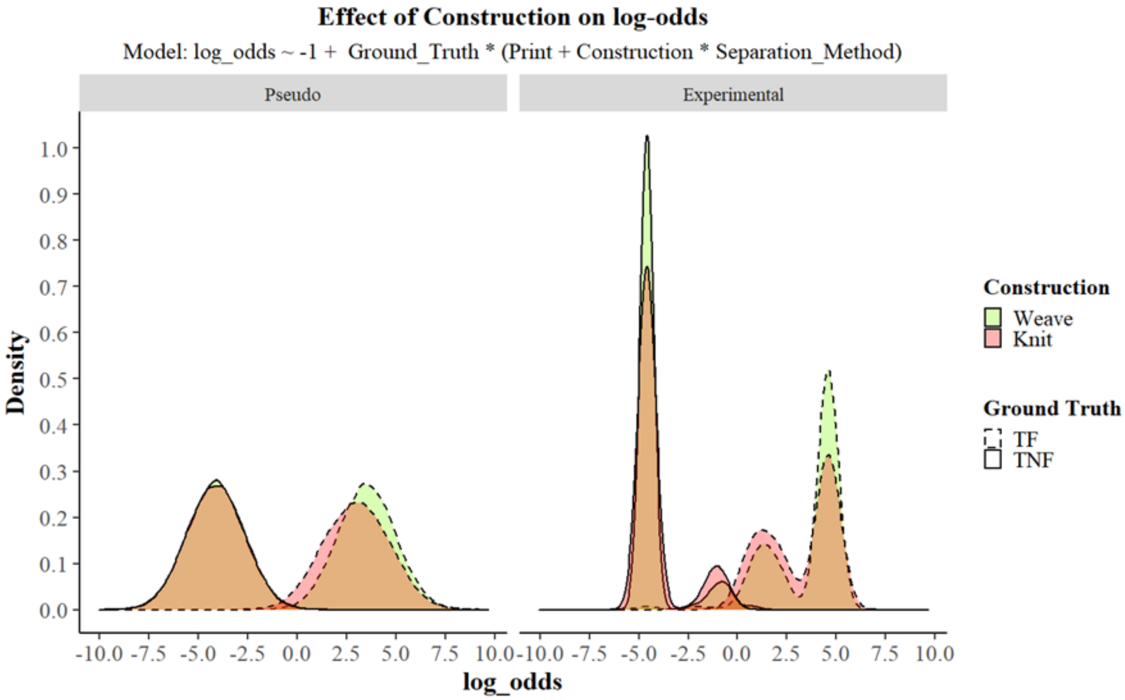


Figure 10. *Counterfactual plot demonstrating the effect of construction (weave or knit) on edge similarity scores. True fits (TF) are presented in dotted lines, while true non-fits (TNF) are in solid lines.*

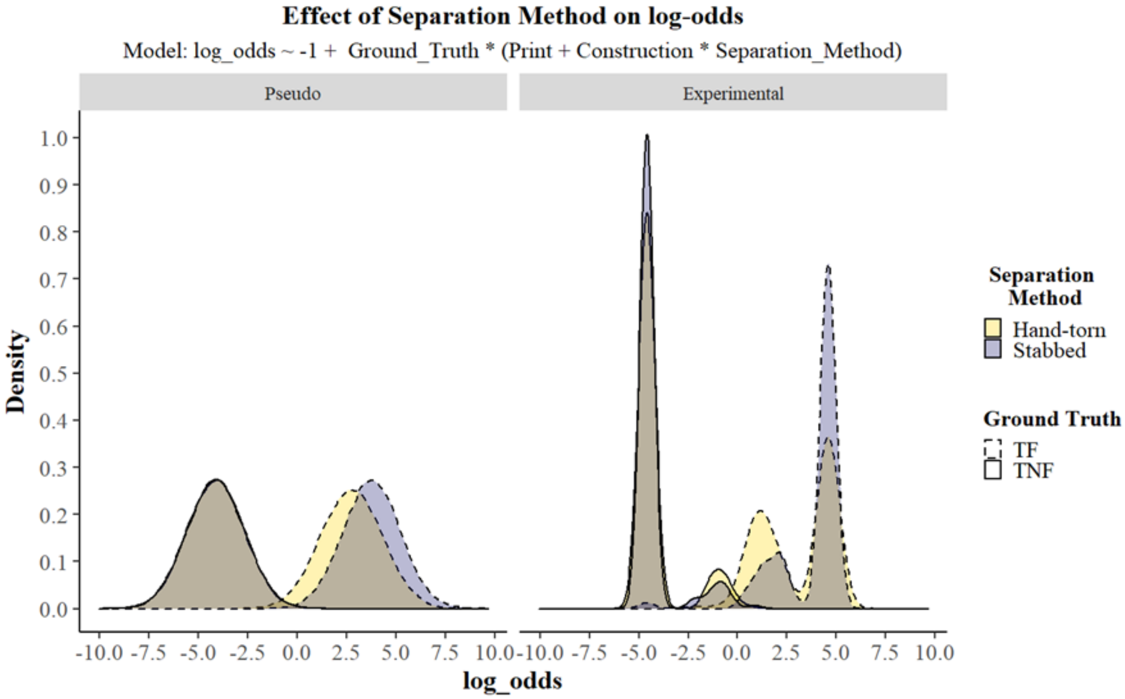


Figure 11. *Counterfactual plot demonstrating the effect of separation method (hand-torn or stabbed) on edge similarity scores. True fits (TF) are presented in dotted lines, while true non-fits (TNF) are in solid lines.*

These plots show a minimal effect on the edge similarity scores by the construction or separation method. As seen in Figure 10, although more true-fitting woven fabrics resulted in high similarity scores compared to the knit fabrics, the densities overlapped so that the overall effect of construction on the scores was negligible. This means that knit fabric did not contribute to considerably lower scores than woven fabric. Figure 11 shows a similar result for the separation method. While more true-fitting stabbed pairs received high edge similarity scores compared to true-fitting hand-torn pairs, the densities for the two separation methods overlap, which indicates that there is not a substantial difference in the effects of the two separation methods on the similarity scores. It is worth noting that the model does not fully account for all potential interactions of the factors, as the experimental data shows multi-modal distributions that are not observed in the simulated data. However, the data indicate that generalizations can be made regarding the ESS scores, regardless of separation method, construction, and design. For instance, in general, scores lower than 20 were indicative of non-fits, while scores greater than 80 were indicative of high-quality fits, with few exceptions. Likewise, the parameter plots, seen in Figure 12, describe the magnitude of the coefficient values for the model, indicating that the magnitude of these values is relatively low across the different factors.

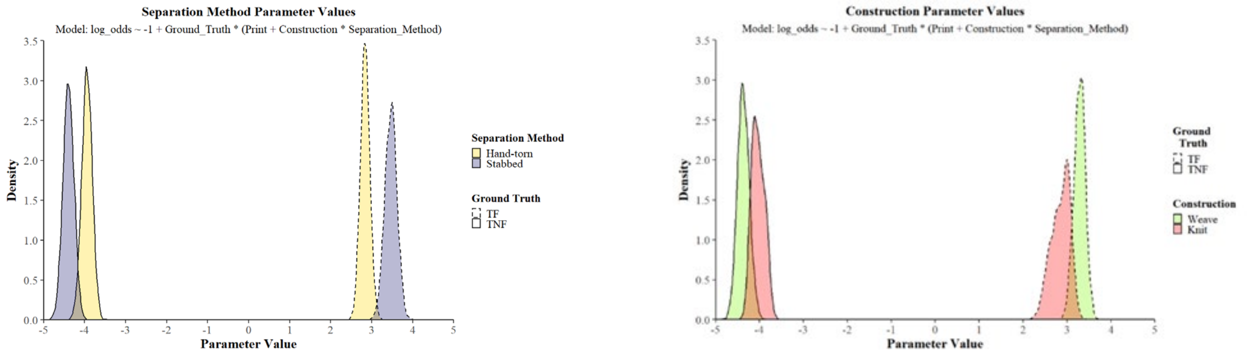


Figure 12. *Parameter plots illustrating the effect of the separation method (left), and construction (right) on the magnitude of the parameters*

Finally, score-based likelihood ratios (SLRs) were calculated to evaluate the probability of observing a specific edge similarity score, given that a comparison is a fit or a non-fit. The calculated SLR values act as a proxy for the probative value of the evidence when compared to a relevant population. Because the logistic regression model did not demonstrate that any of the tested factors substantially affected the similarity scores, the knit and woven hand-torn and stabbed subsets were combined to evaluate the textile set as a whole. The distribution plot of the logarithmic SLR values is shown in Figure 13. These plots display log SLR values, which simplifies interpretation. Positive log SLR values at a specific edge similarity score indicate that the score provides support for a fit decision, with larger log SLR providing stronger support.

On the other hand, negative values observed at a specific ESS value indicate that the score provides support for a non-fit decision. In this case, Figure 13 shows that scores below 10 support a non-fit, with log SLR ranging from 0 to -3 (SLR 1 to 1000). Scores above 60 provided some support for a fit (log SLR 0 to 1, SLR 1 to 10), but ESS values higher than 80 provide strong support for a fit conclusion (log SLR ranging from ~1 to 2.7, SLR 10 to 500). An edge similarity score of 0 results in a log SLR of approximately -3, which indicates that a score of 0 is about 1000 times more likely to occur for a non-fit than a fit. On the other hand, a score of 100 results in a log SLR of approximately 2.7, which indicates that a score of 100 is about 500 times more likely to occur for a fit than a non-fit. It is important to note that because the dataset is somewhat limited in size, few or no values were observed in experimental data at scores ranging from 10 to 60. Therefore this range of scores is not informative for the SLR approach, as observed by the SLR of zero. The SLR approach will benefit from more samples, and expansion of the collection set is recommended. However, the findings provide proof that the ESS can be used as a proxy to evaluate the probative value of a fit or non-fit.

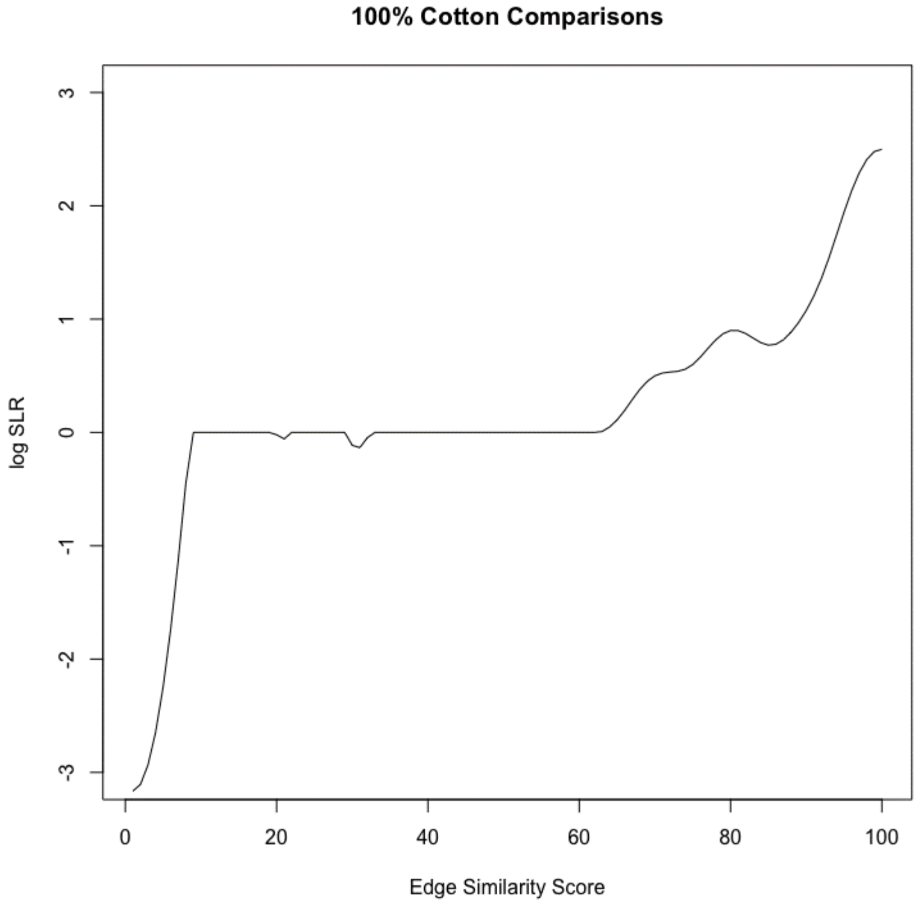


Figure 13. *Plot displaying log score-based likelihood ratios versus the ESS for the 100% cotton textile dataset.*

3.5 CONCLUSIONS

Examiners must consider various factors when conducting physical fit comparisons of textiles. The suitability of the fabric to undergo comparison must be considered first and foremost. Fabrics that deform to an extreme degree may not be suitable for physical fit comparisons because the deformation masks and distorts features critical in identifying a physical fit and negatively impacts accuracy. This study demonstrated that fabrics such as polyester produced highly distorted edges when stabbed or torn, causing unacceptable and misleading error rates. On the other hand, other textiles, such as cotton, demonstrate high accuracy and low misclassification rates. Furthermore, recognizing relevant features, standardized terminology, and reporting criteria are vital for establishing consistency between examiners. The relevant features noted in this study serve as a basis for a more uniform decision-making process. Standardizing terminology and comparison methodology are anticipated to lower intra-examiner variation and increase the repeatability and reproducibility of physical fit comparisons. The systematic nature of this method ensures that comparisons are more objective while enhancing the quality of the peer review process through streamlined communication and documentation of results.

Edge similarity scores for true fits were predominantly in the range of 80 to 100 ESS, while scores for true non-fits clustered around 0 to 10. Overall accuracy for the proposed method was high, ranging from 87% to 100% for different sets, once suitability issues were understood, and relevant features were identified. Over time, performance improved. Training and experience improved accuracy, with rates ranging from 98% to 100% after the examiner had already completed almost 200 comparisons. Intra-examiner variation was also illustrated; when an examiner blindly re-analyzed a set of samples they had previously compared, accuracy improved by 7%. This study highlights the impact of training on standardized protocols and experience on examiner performance. As the examiner became more experienced in physical fit comparisons, including improved recognition of relevant features and knowledge of their impact on the quality of a physical fit, they became increasingly adept at identifying fits and non-fits.

Most misleading rates in the experimental dataset originated from false negative or inconclusive results on true fits. A significant discovery in this study is that not all textiles are suitable for physical fit examinations. Our data indicate that false negatives as high as 63% can be observed in deformable polyesters. This raises a flag, and extended studies encompassing a larger variety of fiber compositions are strongly recommended. Another critical finding in this study is that, although rare, it is possible to observe false positives in fit examinations. Thus, understanding the sources of errors and documenting quality metrics can assist in minimizing misleading evidence rates and provide transparency in the results.

While examiners must consider the composition of the fabric when conducting physical fit comparisons, a logistic regression model demonstrated that other factors, such as separation method and construction of the fabric, did not have a substantial effect on the ESS used as an indicator of the quality of a physical fit. This permits generalizations regarding the dataset. Using score-based likelihood ratios to estimate the ratio of the probability of observing an edge similarity score given that a comparison is a fit or a non-fit, it is evident that ESS scores below 20 provide support for a non-fit decision by an examiner, while scores above 80 provide support for a fit decision. Expanding this dataset to encompass more comparison pairs would strengthen these generalizations and provide an even better understanding of textile physical fits.

Further future work should be devoted to implementing this method through inter-laboratory studies. This would provide valuable feedback from practicing examiners to shape the method in the future. Additionally, further steps could be done to enhance the realism of the textile samples to better represent casework, such as by adding blood to samples in the stabbed subsets.

4. EVALUATION OF THE SCIENTIFIC FOUNDATIONS OF THE INDIVIDUALITY OF THE ORIENTATION OF MICROFIBERS ON PAPER FRACTURE EDGES

4.1 OVERVIEW

This study aimed to assess the occurrence of microfiber alignment along the edges of office paper as means of quantifying the quality of a physical fit in combination with an edge similarity score metric. It is assumed that the orientation and distribution of microfibers on the surface of a sheet of paper are random and therefore provide individuality along the fracture edge when present. For this reason, it is thought that when comparing the edges of two fragments of paper that are true non-fits, there should be no microfibers present that align across the fracture edge between the two samples. Additionally, this study tested the effects caused in the ESS score by differences in the type of paper or printer used to create the sample documents and the separation method (cut or torn).

For this study, two sources of office papers were used that differ in manufacturer and brightness. Paper brightness is measured as a metric of how much the paper reflects blue light. It is measured on a scale of 0-100[53], where brighter paper provides increased contrast with black ink. Both printers used in this study are laser printers that use electrostatic to attract toner particles to a drum that passes over the paper as it feeds through the printer.

A dataset of 160 paper comparison pairs was created using printed and handwritten documents, with the same paragraph of text repeated over the entire page. The handwriting was performed by a person whose handwriting was unfamiliar to both examiners. The dataset was evenly split between scissor-cut and hand-torn samples to assess differences in the separation method. It was hypothesized that scissor-cut pairs would exhibit more instances of microfiber alignment than hand-torn pairs due to the distortion and shearing caused by the hand-tearing process. The dataset was balanced between true fits and true non-fits. One examiner completed the comparisons for the handwritten subset, while a second examiner completed the comparisons for the scissor cut dataset.

The overall accuracy of the method was found to be high for both subsets and separation methods, ranging between 92.5% and 100% for different sets. No false negative misclassifications were observed, but two false positive misclassifications were reported. Both false positive misclassifications were from the printed scissor cut dataset, and both were printed using the same paper and printer. Aligning microfibers were regularly observed in true non-fitting comparisons, though they were generally observed in far fewer numbers than those in true fitting comparisons. Minimal overlap was observed between the number of aligning microfibers documented in true fits and true non-fits. Variance was observed in the number of aligning microfibers that the examiners identified. In general, inter-examiner consistency was good, except that, on average, Examiner 1 generally identified substantially more aligning microfibers in scissor cut pairs than in hand-torn pairs, as compared to Examiner 2.

This study provides a strong foundation for assessing the quality of a physical fit for paper materials. However, variation in results between examiners, while not significantly affecting performance, must be further studied by the expansion of the dataset. This would allow the generalization of these results for a wider population.

4.2 METHODS

4.2.1 Participant Training and Experience

The participants in this study were one graduate and one undergraduate student within our research group. Prior to beginning the comparisons, each participant conducted a thorough review of relevant physical fit literature, which included physical fit studies of various materials, and each had previously been trained on the ESS method for duct tape, including the identification of relevant features. The participants were required to complete an initial set of 40 duct tape comparisons, undergo a review session, and then complete an additional set of 80 duct tape comparisons. The participants' ESS results were compared to previously established consensus scores for each comparison pair to ensure accuracy. The graduate student participant had previously completed over 800 textile physical fit comparisons (Chapter 3), and both students had previously completed 100 postage stamp physical fit comparisons each (Chapter 5). The participants in this study will be referred to as "examiners" for the remainder of this chapter.

4.2.2 Sample preparation

The paper dataset for this study consisted of 160 comparison pairs using multiple separation methods, printing methods, and brands of paper. Figure 14 shows a sampling diagram describing each variable used to design the set.

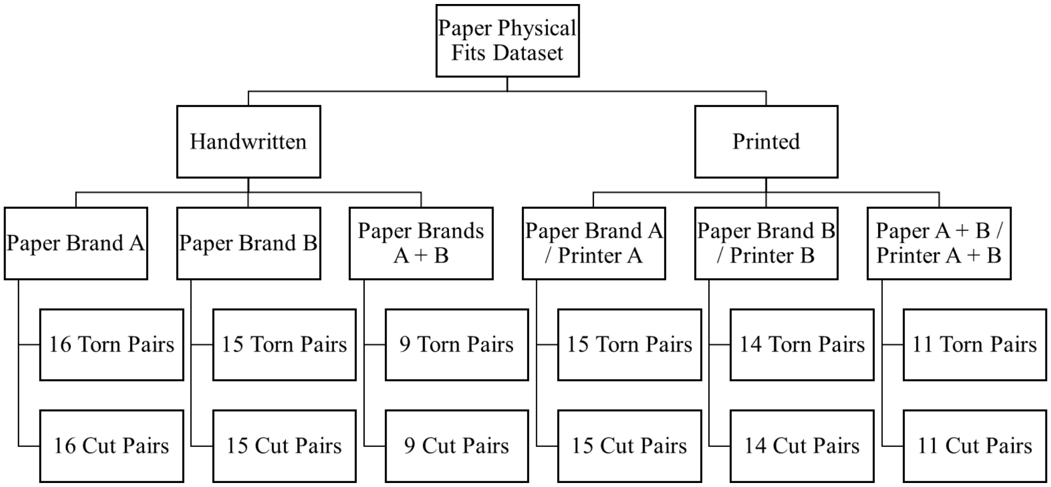


Figure 14. *Diagram of the distribution of paper comparison samples by writing, paper brand, printer type, and separation method. The "Paper Brands A + B" subset refers to a subset of only true non-fitting pairs that are comprised of one fragment that originates from Paper A and the other fragment from Paper B. For the printed subsets, Paper A was only printed from Printer A, and Paper B was only printed from Printer B.*

The same body of text was used to create both the handwritten and printed subsets. A commonly used handwriting exemplar was used, known as the London Letter. This text was selected because it contains every single-digit number and every letter in the alphabet. The London Letter includes the following:

> *Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott Street and then goes to Turin and Rome and will join Colonel Parry and arrive at Athens, Greece, November 27th or December 2nd. Letters there should be addressed King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the 'Y. X.' Express tonight.*

For this study, two brands of standard white office paper were used. Specifications for each product can be seen in Table 11. For the handwritten subset, a single student, who was not involved in conducting the comparisons, handwrote the London Letter repeatedly on eight sheets of paper. The examiners conducting the comparisons were generally unfamiliar with this student's handwriting and did not participate in the experimental design. The printed subset was created by printing a Microsoft Word document containing several repetitions of the London Letter on a single page. Two printers were used in this study to print the documents, and their specifications can be seen in Table 12. Printer A was used only to print Brand A paper, and Printer B was used to print Brand B paper.

Table 11. Specifications for the two brands of paper used in this study

	Brand	Manufacturing Location	Distribution Location	Paper Type	Brightness	Weight (lbs)	Size
Paper A	Tru Red	United States	Massachusetts, USA	Copy	92	20	8.5" x 11"
Paper B	Staples	Canada	Massachusetts, USA	Multipurpose	96	20	8.5" x 11"

Table 12. *Specification for the two printers used in this study*

	Brand	Printer Type	Ink Type	Cartridge
Printer A	HP	Laser	Toner	HP 305X
Printer B	HP	Laser	Toner	HP 410X

Following this, a researcher not involved in the comparison process subdivided each sheet of paper into a grid. Each cell in the grid was given a unique identification number, which was recorded to preserve the ground truth. The grid ensured that each individual sample was of uniform size. An example of this can be seen in appendix II. In this case, the samples were 3 cm x 2 cm, with the comparison edge being the longer of the two sides. True fitting pairs were made from pairs of samples immediately adjacent to each other on a single sheet of paper. Non-fit pairs were made by pairing a sample on one sheet of paper with a sample on another sheet of paper with the same placement in the grid as the first sample's true fitting mate on the first sheet of paper.

To make true non-fitting pairs, the two sheets of paper were overlaid on top of each other so that when they were cut or torn, the edges of the two sheets would be as identical to each other as possible. The cut subset was prepared by using scissors to cut as straight of a line as possible to separate the two samples. The hand-torn samples were prepared similarly so that true non-fitting comparison pairs could not be easily distinguished based solely on significant differences in the edge shape. Additionally, some true non-fitting pairs were made by mixing samples, one edge from Paper A and another from Paper B or Printer A and Printer B, respectively.

4.2.3 Comparison methods

Before comparison, all samples were imaged as pairs using a MiScope® (Zarbeco; Succasunna, NJ) portable microscope. Ten images were captured per pair. Five images were captured using transmitted white light, while another five images were made using transmitted white light and reflected UV light to best visualize microfibers embedded within the paper. Each image captured 6 mm of the 30 mm length of the comparison edge.

Every examiner in this study used a EZ4 stereomicroscope (Leica Microsystems; Deerfield, IL). Examiners were free to use any level of magnification achievable using the stereomicroscope, though 35X magnification was found to be ideal for this study. Transmitted light was used to illuminate the samples and provide optimal visualization of microfibers. Pairs were placed on the stage under a glass slide, which was marked with tick marks every 3 mm to visualize ten comparison bins. When conducting the analysis, the examiner would focus their attention on a single bin comparison at a time, rendering a decision, either fit, non-fit, or inconclusive, for that individual bin comparison before moving on to the next bin. Using a documentation template, which can be seen in appendix II, the examiner noted such features as edge alignment, print/writing alignment, and the presence of any distortion or tearing. When the examiner observed the presence of an aligning microfiber, this was also documented in the template. The aligning microfiber was also traced on the image of the comparison bin using an iPad Air (Apple Inc.; Cupertino, CA), an Apple Pencil, and the Apple Photos annotation feature (see example in Figure 15).

As previously stated, each bin comparison would receive its own individual decision, noted by either a 1, 0.5, or 0. This represents a fit, inconclusive, or non-fit, respectively. These classifiers were then summed for each bin of a comparison and divided by 10 to calculate the Edge Similarity Score for the comparison, which quantifies the quality of the physical fit. The examiner also selects a +/- qualifier to describe their confidence in their decision, with "+" indicating high confidence and "-" indicating low confidence in a fit or non-fit. Typically, high confidence conclusions correspond to an ESS of above 80% for a fit or below 20% for a non-fit.

4.2.4 Data analysis

Performance rates were used to evaluate the overall rigorousness of the method. False-positive rates, false-negative rates, sensitivity, specificity, and overall accuracy were calculated for each of the two examiners. These metrics consider five outcomes of the comparative analysis process as previously explained for textiles (see section 3.2.3). Performance rates are calculated using the number of times these five outcomes occur in a dataset, and the equations used for these rates can be seen in Table 1.

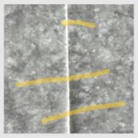
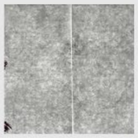
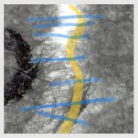
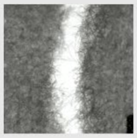
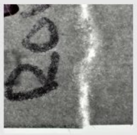
Boxplots and frequency distributions were used to analyze the number of microfibers assigned to both true fits and true non-fits. Box plots graphically display the spread of the data through quartiles, which subdivide the data based on five data points: the median, maximum, minimum, first quartile, and third quartile. These plots can visualize the spread of the number of aligning microfibers assigned to true fits and true non-fits in the data set. Frequency distributions provide a similar view of the data. They are helpful in exploring potential overlap in the data if, for example, twelve aligning microfibers were identified in both a true non-fit comparison and a true fit comparison. These plots help investigate if a threshold could be established for fit/non-fit classification based on the number of aligning microfibers identified by an examiner.

4.3 RESULTS AND DISCUSSION

4.3.1 Standardization of Relevant Features

The first step of developing a systematic method for assessing the quality of a physical fit for paper is the establishment of relevant features that influence the examiners' decision-making process. Standardization of terminology and descriptors for these features makes review and discussion between examiners more straightforward and transparent, as all examiners should recognize the same features and use the same terminology to reference them. Before beginning the comparisons for this study, the examiners created small preliminary sets of paper pairs to identify features that were determined to be most common and distinctive. Three examiners independently evaluated the dataset and six major features were selected after discussion of their observations, which can be seen in Table 13.

Table 13. *Relevant features for the identification of a physical fit for paper materials*

	Embedded Microfiber Alignment	**Extraneous Microfiber Alignment**	**Letter Alignment - Printed**	**Letter Alignment - Handwritten**	**Feathering Alignment**	**Edge Shape Alignment**
Description	Alignment of microfibers embedded within the paper surface	Alignment of microfibers that are protruding out of the fracture edge	Alignment of printed letters or words across the fracture edge	Alignment of handwritten letters or words across the fracture edge	Delamination of paper causes separation of layers within the sheet at the fracture edge	Alignment of overall edge shape and morphology at the fracture edge between two samples
Scissor-Cut		Not Present			Not Present	
Hand-Torn						

The presence and influence of these features can vary based on the separation method or lettering method. Embedded microfiber alignment refers to the alignment across the separated edges of paper microfibers. This microfiber alignment (MFA) is observed at 10-20x magnification. Generally, it occurs more frequently in scissor-cut pairs than hand-torn pairs, due to distortion of the fibers when tearing the paper. On the other hand, the opposite is true for extraneous microfiber alignment, which are fibers that protrude out of the fracture edge and continue to the matching item's edge. These occur more frequently in hand-torn pairs due to the ripping and pulling of the paper when it separates.

The presence of letter alignment provides distinctive features, regardless of the separation method, though the interpretation of this feature may vary depending on the lettering method. In handwritten pairs, letter alignment is more distinctive, as variations in human handwriting, even within writing from the same individual and pen, can be notable. This makes the discovery of aligning letters carry more weight for handwritten documents in comparison to printed documents. The latter contains more reproducible text that can align across a fractured edge even in true non-fitting pairs if the edge shape also aligns, provided the text is of the same font design and size.

Feathering alignment only occurs in hand-torn pairs and occurs when the sheet of paper begins to delaminate at the fracture site, exposing the inner layers of the sheet of paper, which can align on top of the other when joined together across the edge of a true-fitting hand-torn pair. Finally, the edge shape alignment refers to the realignment of the contours of the fracture when the two items are joined together. The contours can be straight, angled/wavy, puzzle-like, or a combination of the above, and observations include alignment of surface features along the edges, such as stains, inclusions, and texture, to mention some.

The designed reporting template provides methodical documentation of any of these features per comparison bin, as well as observations that helped the examiner to form an opinion (see Appendix II for an example of the template). The standardization of these features, both in regards to description criteria and terminology, assists with consistent reporting between examiners. This minimizes variation in the examiners' decision-making process, marking an essential first step in standardizing documentation of these comparisons. In addition to the reporting template, images of the comparison bins allow the examiner to highlight the spatial location of the aligning microfibers and produce a permanent record of the observations (see Appendix II).

4.3.2 Handwritten Paper Comparison Set

Two methods of lettering were used in this study to determine if there would be any effect on an examiner's performance due to differences between handwritten and typed letters. It was hypothesized that the typed comparison pairs might produce more misclassifications compared to the handwritten comparison pairs due to variability in an individual's handwriting that is not seen in more reproducible typed documents. Results for the handwritten subset were promising, exhibiting accuracies of 100% and 97.5% for the cut and torn comparisons, respectively, using the edge similarity score method. Table 14 displays the performance rates for the handwritten subset.

Table 14. *Performance rates for the handwritten paper subset using the edge similarity score method*

Handwritten Paper Subset	Reported Fit	Reported Non-Fit	Reported Inconclusive	Total Comparisons
	Cut Subset			
True Fit	19 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	19
True Non-Fit	0 (0% False Positive)	21 (100% True Negative)	0 (0% Inconclusive)	21
Accuracy	100%			
	Torn Subset			
True Fit	19 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	19
True Non-Fit	0 (0% False Positive)	20 (95% True Negative)	1 (5% Inconclusive)	21
Accuracy	97.5%			

No misclassifications were reported for the scissor-cut comparisons, though one misclassification was reported for the hand-torn comparisons. This comparison pair was a true non-fit, but it was classified as "inconclusive" by the examiner. The pair received an edge similarity score of 50%, and 13 aligning microfibers were documented. Both of the samples in this true non-fitting pair were made with Paper A. In their comparison notes, the examiner observed that the general edge morphology of the pair was strikingly similar, especially in regions that did not display any letter alignment. Figure 15 shows selected images of this comparison, complete with the examiner's annotations. Figure 15A shows Bins 3 and 4, where the examiner classified Bin 3 as inconclusive and Bin 4 as a non-fit. The examiner noted that the stem of the "n" in Bin 4 seemed to be interrupted, which led to the non-fit decision for that bin. Bin 3 was reported as inconclusive due to a similar reason, though the interruption of the leg of the "n" was less clear. Figure 15B shows Bins 5 and 6, where the examiner reported that both bins were a fit. In these bins, there are no instances of letter alignment, and the general edge shape corresponds between the two samples with some feathering alignment observed. The examiner documented three aligning microfibers in Bin 5 and one aligning microfiber in Bin 6. Figure 15C shows Bins 9 and 10, which were both classified as non-fits by the examiner. Microfiber alignment was not observed in either comparison bin, and there is a somehow notable letter misalignment in Bin 10. The spur of the "s" does not align with the spine of the letter. Additionally, the angle between the stress of the "e" and its finial does not appear to align completely.

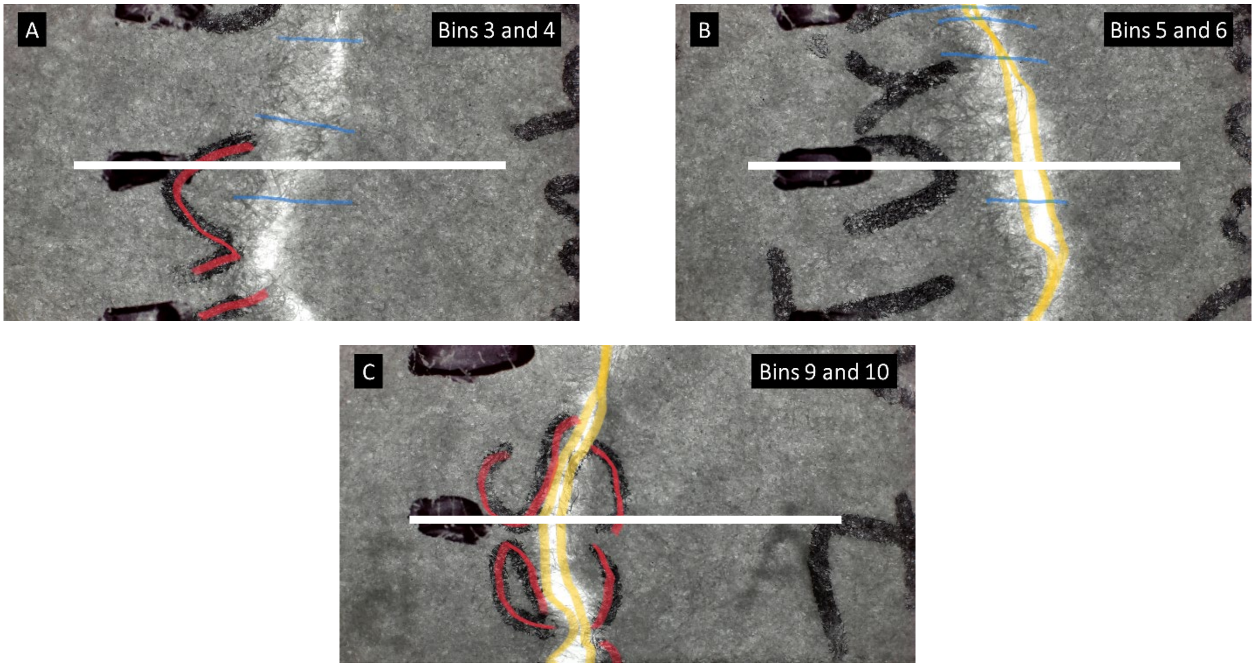


Figure 15. *Comparison images of a true non-fitting comparison pair reported as inconclusive by the examiner. Images are annotated by the examiner during the comparison. Blue highlights indicate aligning microfibers. Yellow highlights trace the comparison edge. Red highlights indicate letter misalignment. The horizontal white bar in the center of each image divides the two bins.*

Interestingly, the examiner observed a significant difference in the number of aligning microfibers between true fitting scissor-cut pairs and true fitting hand-torn pairs. The average number of aligning microfibers observed in a true fitting scissor-cut comparison was 60, while an average of 34 aligning microfibers was identified in true fitting hand-torn pairs. While fewer aligning microfibers were observed in hand-torn pairs, the overall performance did not significantly suffer. The lower number of aligning microfibers observed in hand-torn pairs results from the difference in edge morphology between the two subsets. Hand-torn paper exhibits significantly more features than scissor-cut edges. The hand-torn edges display fraying and feathering, which makes the identification of aligning microfibers much more difficult. Scissor-cut edges are much straighter and do not display the damage imparted by tearing that the hand-torn subset does.

Additionally, the hand-tearing edges commonly display fibers that are pulled out from the edge due to the tear, which hinders the examiner's ability to observe microfiber alignment. This does not happen in the scissor-cut subset, whose edges are much more similar to postage stamp edges. The frequency distribution of aligning microfibers for the scissor-cut and hand-torn subsets can be seen in Figures 16 and 17, respectively. For both cases, there is a clear separation between the number of aligning microfibers observed in true fits and true non-fits. While the distribution is relatively different for the two subsets regarding true fits, both have similar distributions for true non-fits (i.e., less than 10-19 instances of microfiber alignment per 3cm edge). For most true non-fitting comparison pairs, less than ten aligning microfibers were observed. No true non-fitting comparison pair exhibited more than thirteen aligning microfibers; the highest number of aligning microfibers observed in a pair correctly classified as a non-fit was twelve.

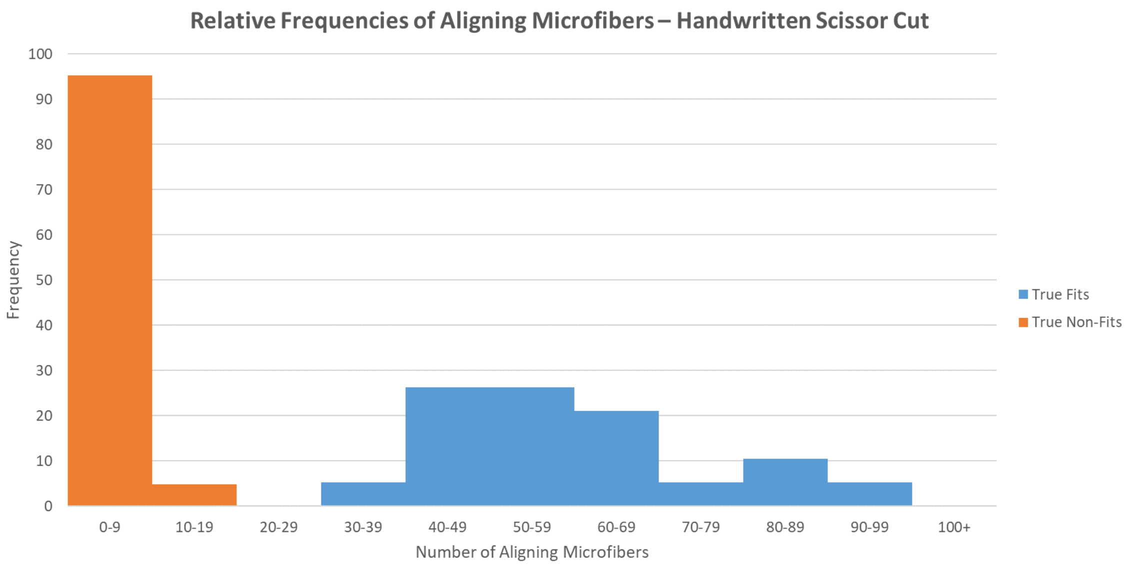


Figure 16. *Frequency distribution of the number of aligning microfibers observed in true fits and true non-fits for the Handwritten Scissor-cut subset*

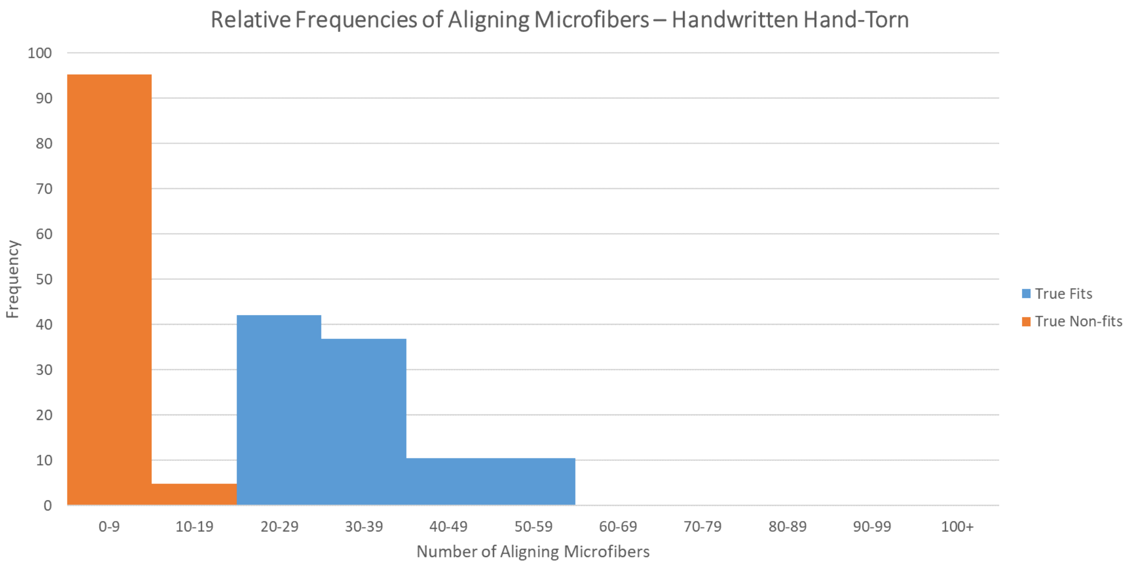


Figure 17. *Frequency distribution of the number of aligning microfibers observed in true fits and true non-fits for the Handwritten Hand-torn subset*

Because there is no overlap between true fits and true non-fits in the number of aligning microfiber observations, there is evidence to suggest that a classification threshold could be implemented to classify comparisons objectively as either a fit or a non-fit based on the number of aligning microfibers observed by the examiner per comparison length. While this threshold would not be a single aligning microfiber, as previously assumed for true fits, this would still permit aligning microfibers to be used as a quantifier for the quality of a physical fit, in addition to the ESS. For this subset of data, a conservative classification threshold of 20 aligning microfibers could be implemented that classifies all comparisons with at least 20 aligning microfibers documented as a fit, with more confidence on a fit result the larger the number of microfiber alignments found. Table 15 displays the performance rates for the dataset, which have been recalculated to reflect the implementation of a 20-aligning microfiber threshold. Using this threshold, the accuracy for the scissor-cut comparison pairs remains at 100%, while the accuracy for the torn comparison pairs is improved to 100%. The true non-fitting pair that was previously reported as inconclusive by the examiner is now correctly classified due to the low number of aligning microfibers observed by the examiner.

Table 15. *Performance rates for the handwritten paper subset reflecting the implementation of a threshold of 20 aligning microfibers (MFA).*

20 Aligning Microfiber Threshold	Reported Fit	Reported Non-Fit	Reported Inconclusive	Total Comparisons
	Cut Subset			
True Fit	19 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	19
True Non-Fit	0 (0% False Positive)	21 (100% True Negative)	0 (0% Inconclusive)	21
Accuracy	100%			
	Torn Subset			
True Fit	19 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	19
True Non-Fit	0 (0% False Positive)	21 (100% True Negative)	0 (0% Inconclusive)	21
Accuracy	100%			

Performance rates were calculated for sixty individual thresholds (1-60 aligning microfibers) to evaluate the best possible threshold for this dataset. The false positive and false negative rates were plotted against each other, similar to a Detection Error Tradeoff (DET) graph. The plot for the hand-written scissor cut dataset can be seen in Figure 18, while the plot for the hand-torn subset can be seen in Figure 19. The most interesting zone of these plots is the point or series of points where the two rates intersect. This point or region represents the threshold(s) with the lowest combination of false negative and false positive rates and therefore the best performance. The plots indicate that the optimal thresholds for the handwritten subset are 13-39 aligning microfibers for the scissor-cut subset and 14-21 aligning microfibers for the hand-torn subset. A range is provided because the performance is equal for every threshold in that range. This data corroborates the selection of 20 aligning microfibers as an acceptable threshold for the handwritten dataset.

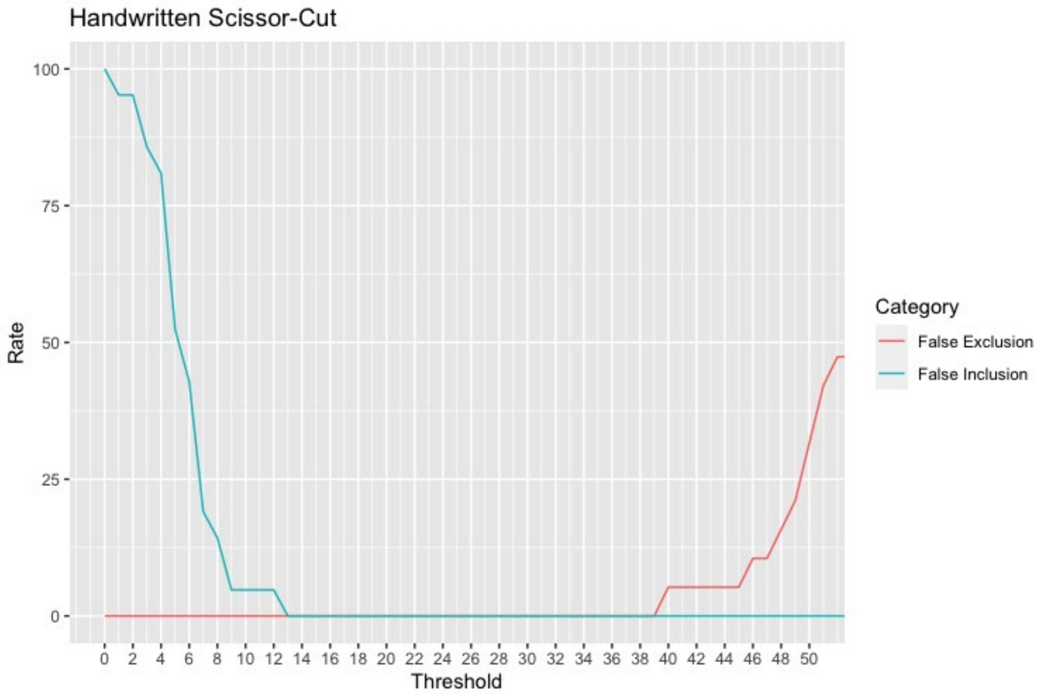


Figure 18. *Plot of false positive rate and false negative rate versus the MFA threshold for the handwritten scissor-cut subset for microfiber alignment thresholds from 1 to 50*

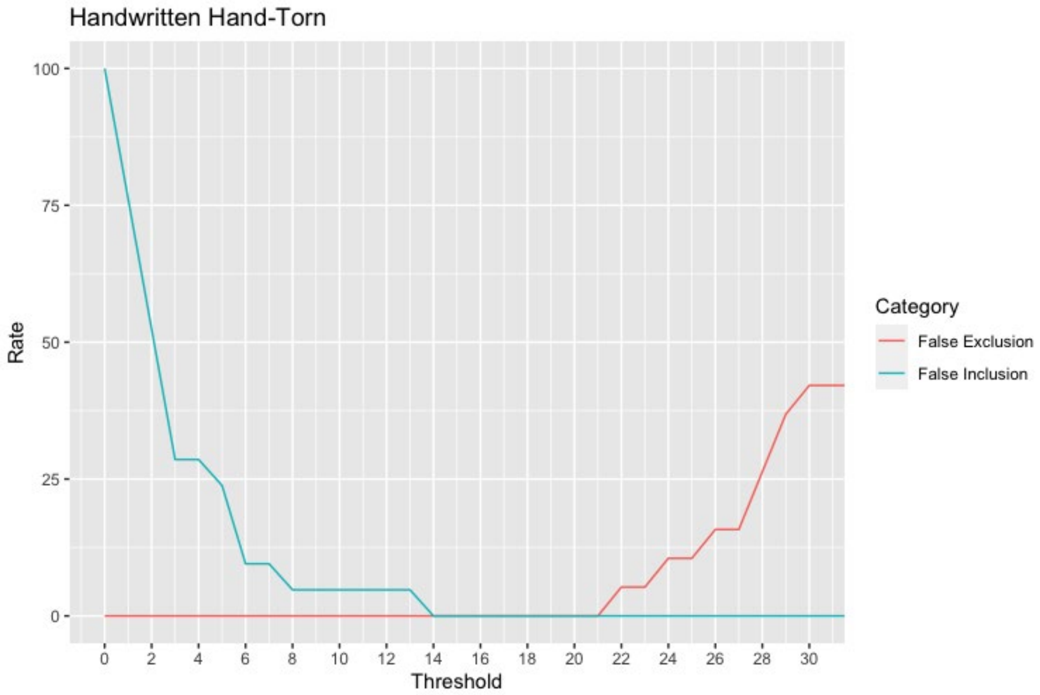


Figure 19. *Plot of false positive rate and false negative rate versus the MFA threshold for the handwritten hand-torn subset for microfiber alignment thresholds from 1 to 30*

Receiver-Operator Characteristic (ROC) curves were made to further investigate the performance of this method for the handwritten dataset. A ROC curve displays the false positive rate on the x-axis against the true positive rate on the y-axis. An ideal ROC curve is “Γ”-shaped, indicating that the metric is effective at identifying, in this research, when two samples originated from the same source. This also indicates 100% specificity. On the other hand, a ROC curve can also be diagonally-shaped, which would suggest that the metric is useless at discriminating between a “match” and a “non-match” and is, therefore, equally likely to conclude that a given comparison results in a “match” or a “non-match.” The method's accuracy can also be determined using a ROC curve by calculating the area underneath the curve, or AUC. The ROC curves for the scissor-cut and hand-torn subsets can be seen in Figure 20.

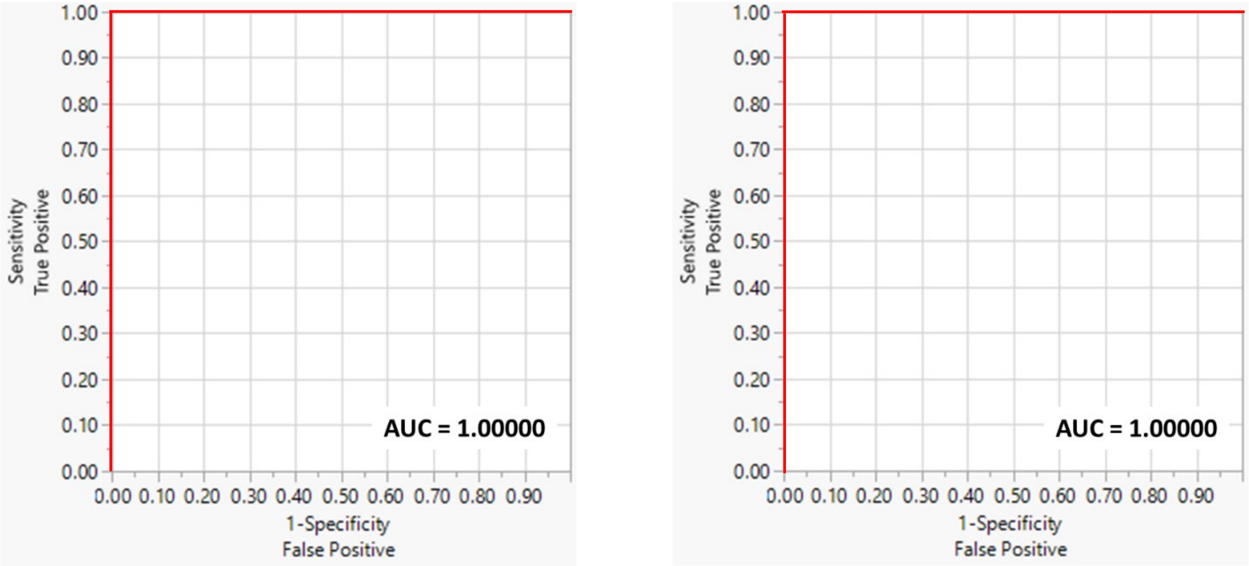


Figure 20. *ROC curve for the scissor-cut (left) and hand-torn (right) comparison pairs of the handwritten subset, showing the examiner classification based on ESS scores*

Both ROC curves have AUC values of 1.0, indicating that the method effectively differentiates between fits and non-fits. As previously stated, there are no false positives observed in this subset. Therefore, a classic, theoretically “perfect” ROC curve is generated. However, this is a relatively small dataset, so future work should be done to expand the sample size and gain an improved understanding of microfiber alignment in office paper.

4.3.3 Printed Paper Comparison Set

The results of the printed paper comparison set supported the hypothesis that typed samples would produce more misclassifications than handwritten samples due to the uniformity in letterform; accuracy for this set slightly fell compared to the previous set. Accuracy for the scissor-cut subset was 92.5%, while accuracy for the hand-torn subset was 100%. A complete summary of performance rates can be seen in Table 16.

Table 16. Performance rates for the printed paper subset using the edge similarity score method

Printed Paper Subset	Reported Fit	Reported Non-Fit	Reported Inconclusive	Total Comparisons
	Cut Subset			
True Fit	18 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	18
True Non-Fit	2 (9% False Positive)	19 (86% True Negative)	1 (5% Inconclusive)	22
Accuracy	92.5%			
	Torn Subset			
True Fit	18 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	18
True Non-Fit	0 (0% False Positive)	22 (100% True Negative)	0 (0% Inconclusive)	22
Accuracy	100%			

Of note is the presence of two false positive misclassifications among the cut subset (9% false positive). While it would not usually be surprising to see a false positive in this case due to the care taken in the experimental design to make every individual sample as uniformly similar as possible with regard to edge shape and letter alignment, the number of microfibers in alignment that the examiner observed was certainly remarkable. For one of the false positive misidentifications, the examiner documented 59 aligning microfibers, and 63 aligning microfibers were observed for the second comparison. To better understand the relationship between these numbers and the number of aligning microfibers that were observed for other true fits and true non-fits in the dataset, frequency distributions were made, which can be seen in Figures 21 and 22, which show the distributions for the scissor cut and hand-torn sets, respectively.

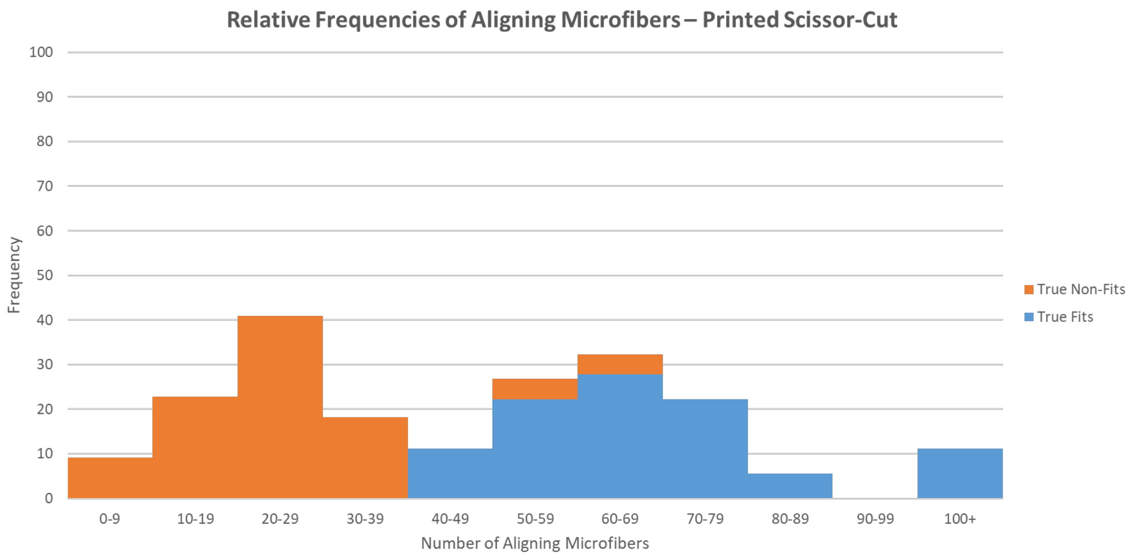


Figure 21. *Frequency distribution for the number of aligning microfibers observed in the printed scissor-cut dataset*

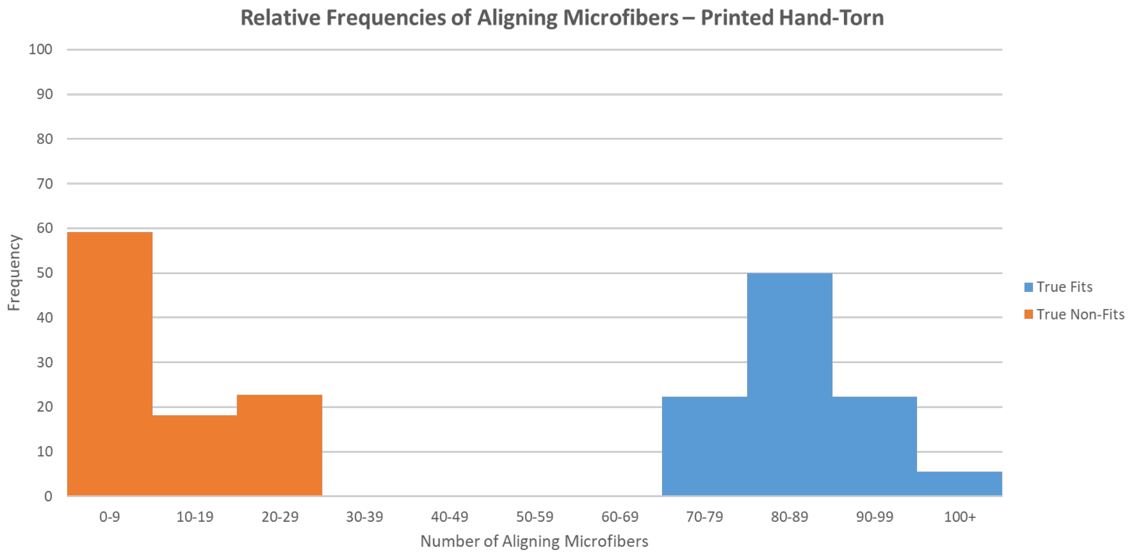


Figure 22. *Frequency distribution for the number of aligning microfibers observed in the printed hand-torn dataset*

While the frequency distribution for the hand-torn subset illustrates a very large gap between the number of aligning microfibers assigned to true non-fits and those assigned to true fits, there is some overlap between true fits and true non-fits for the scissor cut subset. The separation between the fits and non-fits in these types of examinations and metrics is ideal because it shows that the method is adept at differentiating between the two classes (i.e., fits vs. non-fits). Suppose this separation is observed in large enough datasets to generalize the population. In that case, it supports the implementation of a classification threshold to objectively classify comparisons based on the number of aligning microfibers observed by the examiner. Therefore, this study serves as a proof of principle but expanding with larger datasets and paper types is recommended.

As shown in Figure 21, it does become problematic when the overlap between the classes is observed. In this case, the overlap is caused by the two false positive misclassifications in the scissor cut subset. Both true non-fitting pairs were printed using Paper A with Printer A. Images of these comparisons can be seen in Figures 23 and 24.

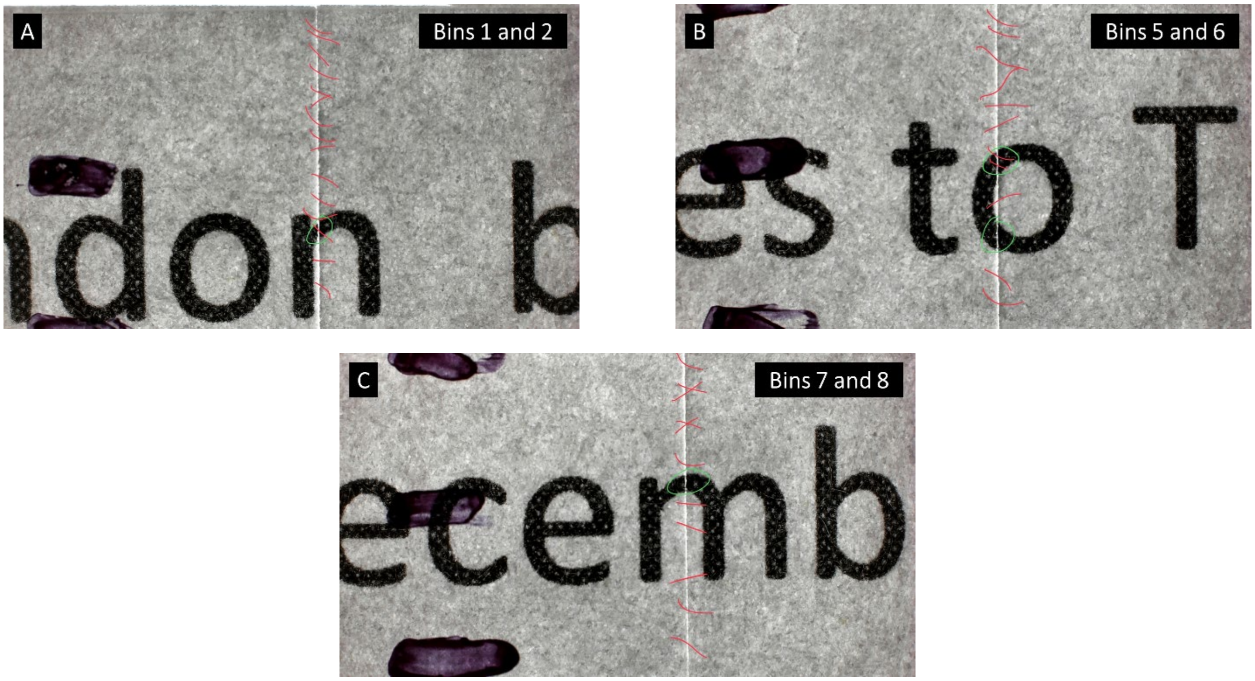


Figure 23. *Comparison images of a true non-fitting comparison pair (P_SC_1196 and P_SC_2240) reported as "fit" by the examiner. Images are annotated by the examiner during the comparison. Red highlights indicate aligning microfibers. Green circles indicate letter alignment.*

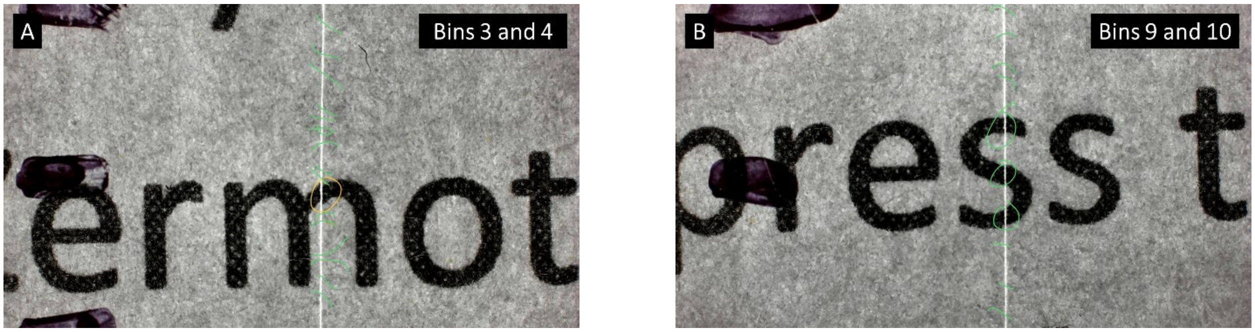


Figure 24. *Comparison images of a true non-fitting comparison pair (P_SC_1344 and P_SC_1259) reported as "fit" by the examiner. Images are annotated by the examiner during the comparison. Green highlights indicate aligning microfibers. Green circles indicate letter alignment. Yellow circles indicate uncertainty regarding letter alignment.*

In Figure 23, various comparison bins can be seen that show the multitude of aligning microfibers that were observed along the comparison edge of this pair, highlighted in red. In total, 59 aligning microfibers were documented for this pair, some of which can be seen in the images. The examiner assigned an edge similarity score of 95% to this pair. Several of these fibers have added evidentiary weight because they cross over with other aligning fibers, forming "X" or "Y"-like shapes across the fractured edges. Examples of this kind of alignment can be seen in Figures 23B and 23C. This alignment provides additional confidence in a physical fit to the examiner due to the assumed unlikelihood of two or more microfibers crossing over each other and aligning across the edge of a true non-fitting comparison pair by chance in non-fitting pairs. That assumption was not met in this case. When viewing the letter alignment for this comparison, the alignment of the "n" in Figure 23A and the "o" in Figure 23B is very good. They do not provide any immediate indications of the possibility of a non-fit. The alignment of the "m" in Figure 23C is also interesting. In this case, the leftmost shoulder, directly connected to the stem, appears slightly elongated compared to the right shoulder. This causes a slight widening of the aperture under the shoulder. However, this was not noted by the examiner at the time of the comparison. A second examiner blindly re-examined this pair due to the unique nature of observing so many aligning microfibers in a true non-fit. The second examiner reached the same conclusion of a fit and observed a similar number of aligning microfibers as the first examiner. From a design point of view, this was a "worst-case scenario" pair where two printed pages were placed on top of each other, and the printed text on two different pages was carefully aligned before cutting the edge with a single scissor stroke. Thus, it was not surprising that features such as edge shape alignment and letter alignment were observed and produced a large score. Although the microfiber alignment was surprising, it raised a flag indicating microfiber alignment should not be used as a sole criterion for a fit.

In Figure 24, two images are shown that contain four comparison bins. Several aligning microfibers can be seen in these images, especially in Figure 24A, where the examiner documented ten aligning microfibers in Bin 3 and six in Bin 4. In total, 63 aligning microfibers were documented for this pair, which was assigned an edge similarity score of 90%. In Bin 4, the examiner also noted their uncertainty in the possible alignment of the letter "m." In this case, the second (right) shoulder appears slightly misaligned with the first shoulder. This area is circled in yellow on the image. This is an area that the examiner must be especially cautious with because it is easy to overlook this misaligning feature in favor of the multitude of aligning microfibers that were also observed. Another example of this can be seen in Figure 24B. The beak of the "s" in the center of the image appears slightly misaligned with the spine. This misalignment is so subtle that the examiner did not notice it and specifically noted the "s" as aligning across the comparison edge. These instances highlight the need for examiners to consider every feature, not just the alignment of microfibers when conducting physical fits of paper materials. Even subtle differences can point to the correct decision.

One interesting true non-fitting comparison from the hand-torn subset can be seen in Figure 25. The examiner documented 22 aligning microfibers for this pair and assigned it an edge similarity score of 40%. In Figure 25A, the alignment of the "d" in Bin 4 can be seen, while Figure 25B shows the misalignment of the "o" in Bin 6. Of particular interest is the alignment of the "m" in Bin 8, which can be seen in Figure 25C. While the upper region of the leftmost shoulder appears to align, the lower leg does not align, which emphasizes another area of importance for the examiner to consider. While this pair did receive a relatively high score for a true non-fit, it is important that the examiner recognizes areas of obvious misalignment, such as the "o" in Bin 6 that indicate that the comparison should be considered a non-fit. Another interesting true non-fitting comparison can be seen in Figure 26, which also received an edge similarity score of 40%. While these two samples both received the same score, they reached this score in two different ways. The first pair exhibited several bins considered inconclusive, with a few non-fitting bins and two fitting bins mixed in. On the other hand, there were no inconclusive bins documented for the second comparison pair. The examiner documented four fitting bins followed by six non-fitting bins. The edge similarity score notation for both comparisons can be seen in Figure 27. It should be noted that both comparison pairs were prepared using the same paper, Paper A.

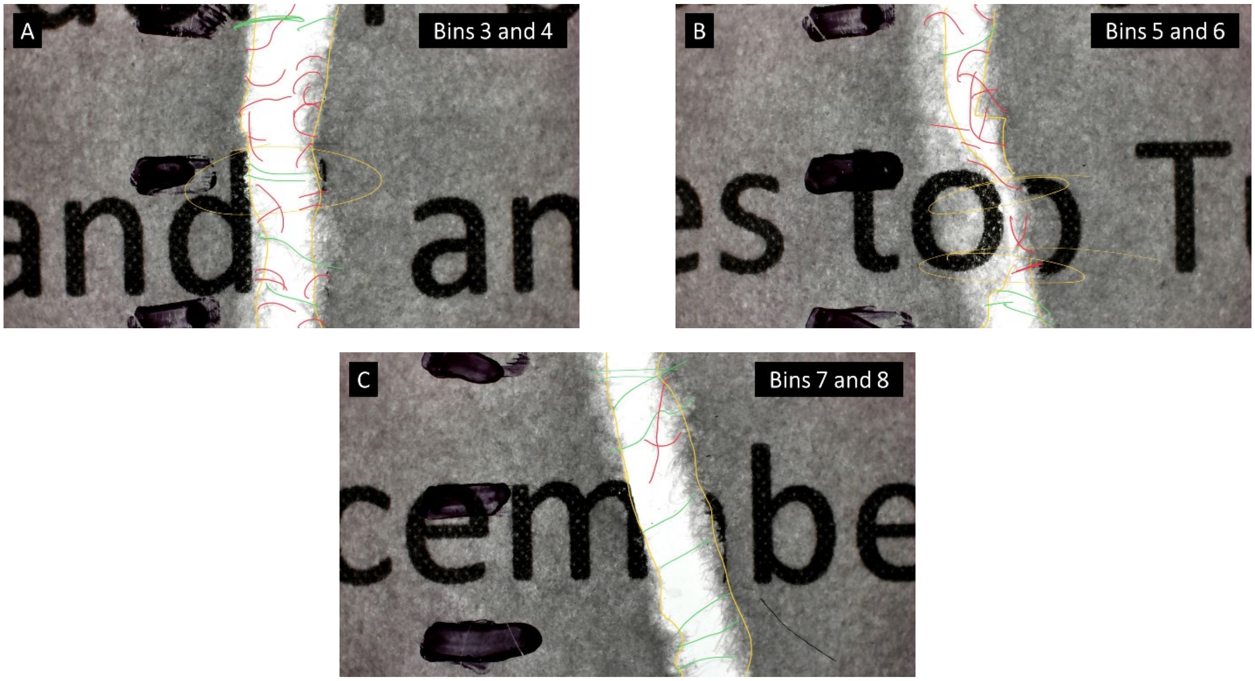


Figure 25. *Comparison images of a true non-fitting comparison pair (P_HT_2159 and P_HT_2200) reported as "inconclusive" by the examiner. Images are annotated by the examiner during the comparison. Green highlights indicate aligning microfibers. Red highlights indicate misaligned microfibers. Yellow circles indicate uncertainty regarding letter alignment.*

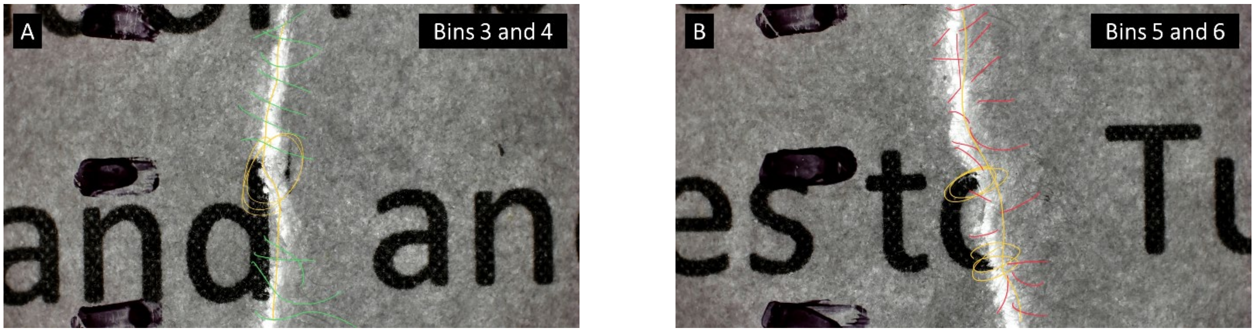
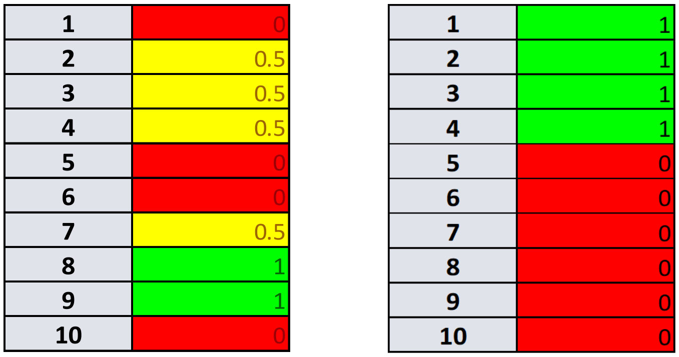


Figure 26. *Comparison images of a true non-fitting comparison pair (P_HT_4699 and P_HT_2455) reported as "inconclusive" by the examiner. Images are annotated by the examiner during the comparison. Green highlights indicate aligning microfibers. Red highlights indicate misaligned microfibers. Yellow circles indicate uncertainty regarding letter alignment.*



1	0
2	0.5
3	0.5
4	0.5
5	0
6	0
7	0.5
8	1
9	1
10	0

1	1
2	1
3	1
4	1
5	0
6	0
7	0
8	0
9	0
10	0

Figure 27. *Edge Similarity Score notations for the printed hand-torn comparison pairs correspond to Figures 25 (Left) and 26 (Right). Here, red cells with the number zero indicate a non-fit bin, green with the number 1 a fit bin, and yellow with the number 0.5 an inconclusive bin.*

For the hand-torn pairs, the frequency distribution shown in Figure 22 emphasizes the presence of wide separation between the number of aligning microfibers attributed to true fits and true non-fits. As previously discussed, the frequency distribution for the scissor cut pairs shows some overlap, though separation does exist except for the two false positive comparisons. For both printed subsets, implementing a classification threshold set at 40 aligning microfibers would improve performance rates, as seen in Table 17.

Table 17. *Performance rates for the printed paper subset reflecting the implementation of a 40 aligning microfiber classification threshold.*

40 Aligning Microfiber Threshold	Reported Fit	Reported Non-Fit	Reported Inconclusive	Total Comparisons
	Cut Subset			
True Fit	18 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	18
True Non-Fit	2 (9% False Positive)	20 (91% True Negative)	0 (0% Inconclusive)	22
Accuracy	95%			
	Torn Subset			
True Fit	18 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	18
True Non-Fit	0 (0% False Positive)	22 (100% True Negative)	0 (0% Inconclusive)	22
Accuracy	100%			

As discussed before, for the handwritten set, performance rates were calculated for sixty individual thresholds (1-60 aligning microfibers) to evaluate the best possible threshold for this dataset statistically. The false positive and false negative rates for the printed comparisons were plotted against each other, which can be seen in Figure 28 for the scissor-cut subset and Figure 29 for the hand-torn subset. The most interesting area of these plots is the point or series of points where the two rates intersect. This point represents the threshold(s) with the lowest combination of false negative and false positive rates and therefore the best performance. For the printed subset, the plots indicate that the optimal thresholds are 37-49 aligning microfibers for the scissor-cut subset and 24-60 aligning microfibers for the hand-torn subset. A range is provided because the performance is equal for every threshold in that range. This data corroborates the selection of 40 aligning microfibers as an acceptable threshold for the printed dataset.

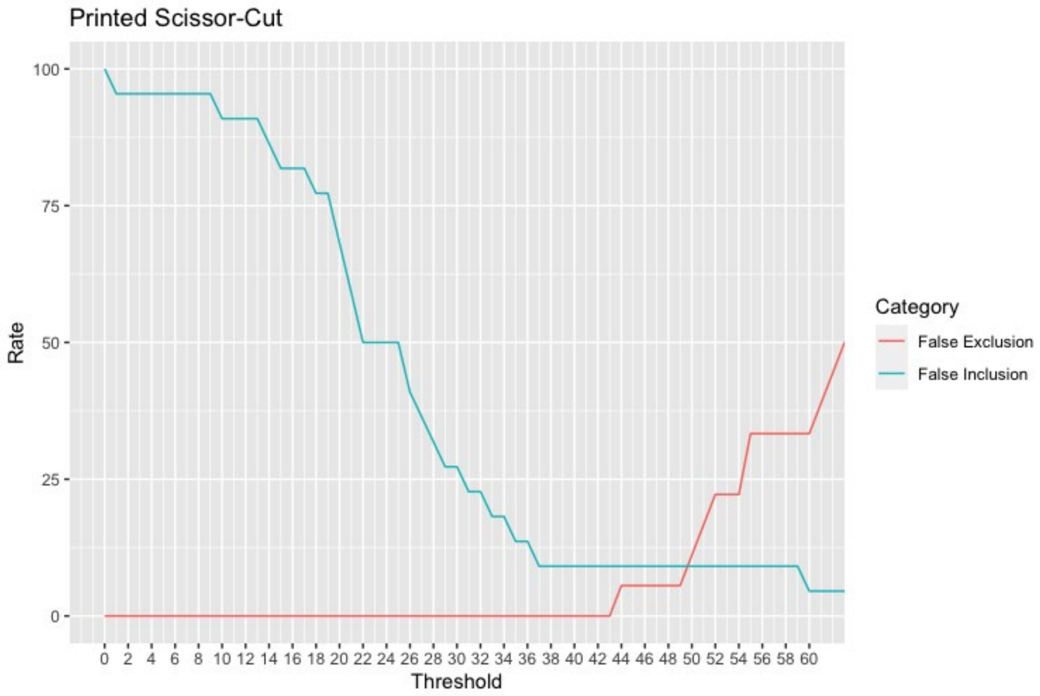


Figure 28. *Plot of false positive rate and false negative rate versus MFA threshold for the printed scissor-cut subset for microfiber alignment thresholds from 1 to 60*

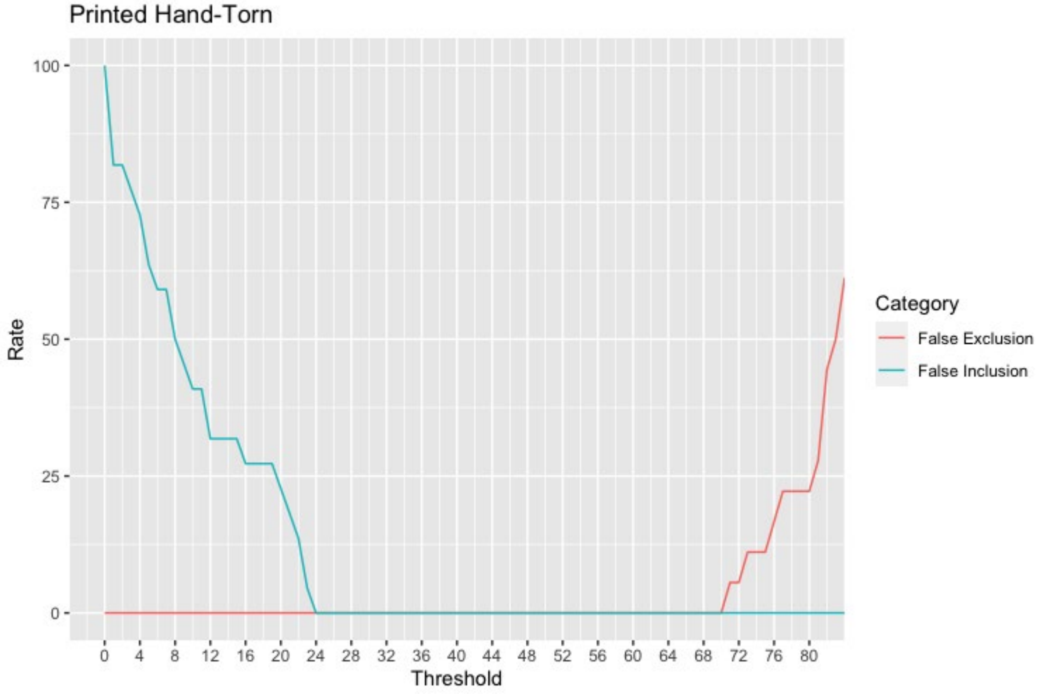


Figure 29. *Plot of false positive rate and false negative rate versus MFA threshold for the printed hand-torn subset for microfiber alignment thresholds from 1 to 80*

The implementation of this threshold does improve performance slightly. The three pairs reported as inconclusive across the set are now all correctly classified as non-fits using the microfiber alignment threshold. However, the two false positive comparisons are still reported as false positives due to the large number of aligning microfibers that were observed for those two pairs. Receiver-Operator Characteristic (ROC) curves were made to further investigate the performance of this method for the printed dataset. The ROC curves for the scissor-cut and hand-torn subsets can be seen in Figure 30. The scissor-cut subset has an AUC of 0.95, while the hand-torn subset has an AUC of 1.0. Both values indicate very good performance by the ESS method at distinguishing between fits and non-fits. However, the scissor-cut subset performance is hampered slightly by the false positive misclassifications. It must be restated that this is a relatively small dataset, so future work should be done to expand the sample size and gain an improved understanding of microfiber alignment in office paper.

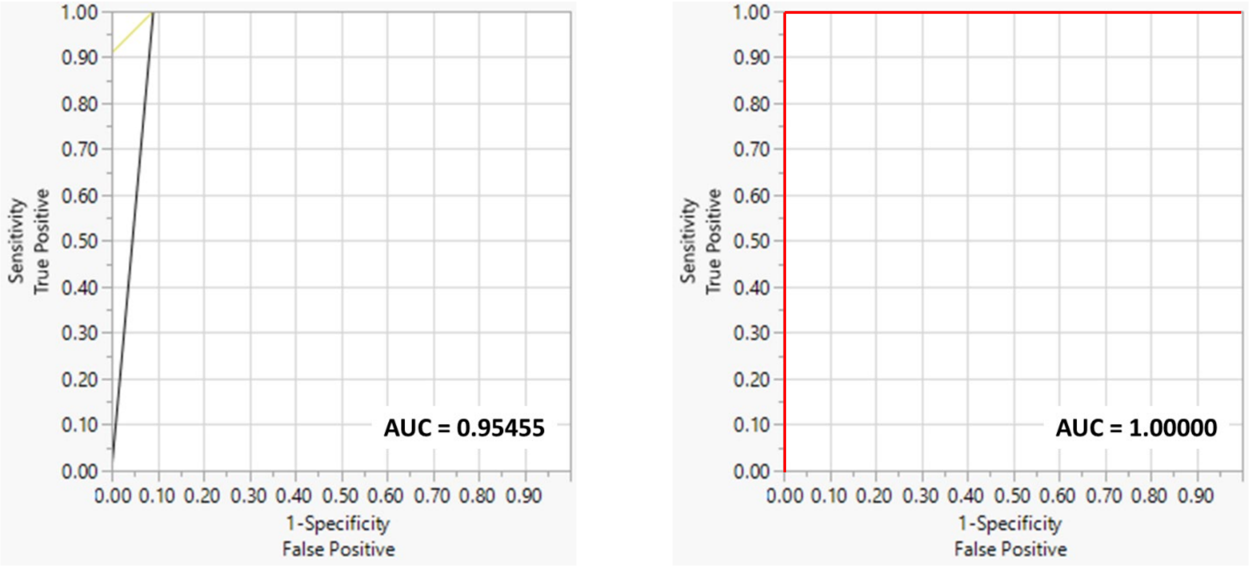


Figure 30. *ROC curves for the scissor cut (left) and hand-torn (right) comparison pairs of the printed subset, showing the examiner classification based on ESS scores*

4.4 CONCLUSIONS

This study developed and evaluated a systematic, quantitative approach for assessing and documenting the quality of a physical fit for paper. A set of 160 comparison pairs of paper edges composed of true fits and true non-fits was used to quantify the occurrence of microfiber alignment on comparison edges and identify relevant features that would affect the alignment occurrence, such as the type of paper or ink entry. The printer model and lettering method were also varied to test how these factors influence the performance in assessing the quality of a physical fit. The ground truth of the comparison pairs was kept blind from the two examiners so that performance rates could be ascertained.

The assumption that no aligning microfibers would be found on the edges of true non-fitting paper comparison pairs was experimentally disproven by this study. True non-fitting comparison pairs were found to exhibit some level of microfiber alignment, which varied based on the separation method. Despite this, there was minimal overlap between the number of aligning microfibers observed in true fits and true non-fits. The edge similarity score metric proved adept at quantifying the quality of a physical fit for paper materials by considering several features that influenced the examiners' opinion of the presence of a fit or non-fit. This metric produced high accuracies ranging between 92.5% to 100%.

The primary feature that proved most indicative of a physical fit was microfiber alignment (MFA), though this feature did vary based on the separation method. Generally, more aligning microfibers were identified in scissor-cut pairs compared to hand-torn pairs. Other important features include letter alignment and edge shape alignment. The latter is usually only relevant in hand-torn pairs, which exhibit variations in edge shape that are not present in scissor-cut pairs; meanwhile, the former can be relevant for both of the separation methods studied. Letter alignment is a viable feature of interest for printed and handwritten documents. However, examiners should be cautious when considering letter alignment for printed documents, as misclassifications could occur due to similarity in typed letters from unrelated documents.

The MFA thresholds established in this study were slightly different for the printed and handwritten subsets. While it is not believed that the type of lettering influences the number of aligning microfibers that an examiner identifies, the data suggests that aligning microfiber observations vary from examiner to examiner, and MFA thresholds should be considered as a "range" that take into consideration this uncertainty. An interesting trend can be observed when comparing the datasets. For the handwritten subset, the examiner identified almost double the number of aligning microfibers for true fitting scissor cut pairs compared to true fitting hand-torn pairs. This emphasizes the need for future studies to expand the sampling size to gain a more thorough understanding of microfiber alignment in paper. An expanded understanding of microfiber alignment would allow the community to determine the possibility of implementing criteria for decision thresholds and the respective error rates.

Additionally, while no significant differences in the number of aligning microfibers were observed between types of paper or printers for strictly comparisons conducted by the same examiner, it should be noted that all misclassifications in this dataset involved Paper A. While this could be coincidental, it could also indicate that the microfibers that are distributed through the surface of Paper A are more uniformly oriented than those in Paper B. This does seem unlikely, though, given the lack of variance in the number of aligning microfibers that were identified when examiner and separation method are held constant. The sample size for this study is small; however, it serves as a proof of principle and sets some criteria and scientific foundations for fractured document comparisons. Expanding the dataset could shed more light on the differences between paper or printer types on examiner performance or microfiber identification.

This study lays the foundation for understanding the occurrence of relevant features and microfiber alignment in paper physical fit comparisons. No substantial differences were observed with regard to the paper type or printer models. Future work should expand this study to provide generalizations regarding the entire population, which would improve upon the foundation outlined in this study.

5. ASSESSING THE VALUE OF MICROFIBER ALIGNMENT BETWEEN POSTAGE STAMP EDGES FOR PHYSICAL FIT COMPARISONS

5.1 OVERVIEW

The purpose of this study involved establishing the scientific foundations of physical fits on postage stamp edges, including the rarity of the orientation of microfibers along consecutive edges. It is assumed that the orientation of microfibers on the surface of a postage stamp is random and therefore provides individuality along the machine-cut fracture edges. Additionally, postage stamps are die-cut in sheets by a machine, reducing the value of the macroscopic edge morphology. Generally, materials that are compared for the presence of a physical fit have been separated by force, which imparts characteristics to the fragments that are not expected to be replicated by chance, which helps identify a physical fit. This is not the case with traditional postage stamps, as the die-cut produces uniform edges that will fit together regardless of whether or not they were oriented side-by-side on the original sheet. Therefore, microfiber alignment on consecutive edges will be used as the primary indicator of a physical fit in this study. This will assess the application of microfiber alignment as a means of quantifying the quality of a physical fit.

A dataset of 100 postage stamp comparison pairs was created using three sheets of traditional adhesive postage stamps. Each stamp was removed from its location on the sheet and placed on an individual acetate square so that the examiner could examine both the upper design side and the lower adhesive side of the stamp. This process was documented to preserve the ground truth of the dataset but was not made available to the examiners until all examinations were completed and reported. After a third party created 100 comparison pairs, two examiners began an analysis of the dataset. The comparisons were conducted by subdividing the edges of the stamps into 13 bins of equal size and conducting bin-by-bin comparisons between a pair of stamps to identify any aligning microfibers. The location within each bin and a short description of each aligning microfiber observed by either examiner were documented. The initial 35 comparisons performed by the examiners were reported, discussed, and re-evaluated as a consensus training set to aid both examiners in standardizing the microfiber alignment assignments and mitigate inter-examiner variations.

The overall accuracy of the method was found to be high, above 91% and reaching 100% in some instances. Despite the common assumption, both examiners observed aligning microfibers in true non-fitting comparison pairs. However, the number of aligning microfibers in true fitting comparison pairs was substantially higher than in true non-fitting comparison pairs. This is critical as it means that microfiber alignment can be used to differentiate between fits and non-fits accurately. A fundamental discovery in this work is that observing a single aligning microfiber in a given comparison does not provide enough support to a fit decision. Therefore, applying a threshold at a specified number of aligning microfibers was explored to increase the objectivity and accuracy of physical fit analysis and provide better support for an examiner's conclusions.

This study provides a novel method for assessing physical fits for postage stamps by using the number of aligning microfibers (MFA) observed on the comparison edge of a pair of stamps as a metric to quantify the quality of a physical fit. This systematic method also provides a strong foundation for transparent peer review. Future work should expand upon the dataset used in this study to generalize these results for a wider population.

5.2 METHODS

5.2.1 Participant Training and Experience

The participants in this study were one graduate and one undergraduate student within our research group. Prior to beginning the comparisons, each participant conducted a thorough review of relevant physical fit literature, which included physical fit studies of various materials, and each had previously been trained on the ESS method for duct tape, including the identification of relevant features. The participants were required to complete an initial set of 40 duct tape comparisons, undergo a review session, and then complete an additional set of 80 duct tape comparisons. The participants' ESS results were compared to previously established consensus scores for each comparison pair to ensure accuracy. Furthermore, the graduate student participant had previously completed over 800 textile physical fit comparisons (Chapter 3). The participants in this study will be referred to as "examiners" for the remainder of this chapter.

5.2.2 Sample preparation

Three sheets of 5¢ postage stamps, containing 20 stamps per sheet, were used to create a dataset of 100 postage stamp comparison pairs. Each postage stamp on a sheet has two to four true positive fits with another stamp on that sheet because the stamps are arranged in a grid pattern. Each stamp was removed from the sheet and adhered onto a pre-cut 2" by 2" acetate square. Each acetate square was labeled with a code representing the sheet that the stamp originated from and the stamp's location on that sheet. Following this, a key was created to document the ground truth, true positive pairs for each stamp. A directional identifier (north/south/east/west) was added to each sample ID to trace the orientation of the true fit comparison pair. For example, sample A-1 would align with sample A-2 in the east-west direction and sample A-6 in the south-north direction (Figure 31). A third party then relabeled each sample using a random number generator and scrambled the comparison pairs to create the 100 comparison pairs for this dataset, balanced between true fits and true non-fits.

5.2.3 Comparison methods

Two examiners independently conducted each comparison by subdividing the length of the comparison edge into thirteen bins of equal size. This number was chosen by consistently dividing the wave pattern along the edge of each stamp into waves of equal size, as seen in Figure 31. The pairs were analyzed using an EZ4 stereomicroscope (Leica Microsystems; Deerfield, IL) with transmitted light for optimal visualization of the microfibers. The examiners conducted bin-by-bin comparisons for each pair by identifying the number of aligning microfibers within each bin. The examiner then classified each pair as either a fit, non-fit, or inconclusive, based primarily on the number of aligning microfibers identified in each bin. Microfiber alignment was documented by noting the location of the microfiber in the bin using three possible locations per bin, in relation to the left-hand sample, from the top: upper crest, trough, and lower crest. The number of microfibers in each location for each bin was documented, and the bins were summed to estimate the total number of aligning microfibers per edge. An example of the documentation template can be seen in appendix III.

Microfiber alignment was the primary feature of interest for these comparisons because the stamps all had the same design and were machine-cut. For physical fit examinations of other materials, such as paper or textiles, the fracture event, whether tearing or cutting, often impart distinctive features that can be used to identify a physical fit. This is not the case with stamps, as they are all separated in a controlled, repeatable manner, and the borders are plain with no design crossing from one edge to another. Other than microfiber alignment, the only identifying feature that can identify a physical fit is damage or any other alteration along the comparison edge observed between the two samples. Nonetheless, this was not the case in this set since these samples were pristine.

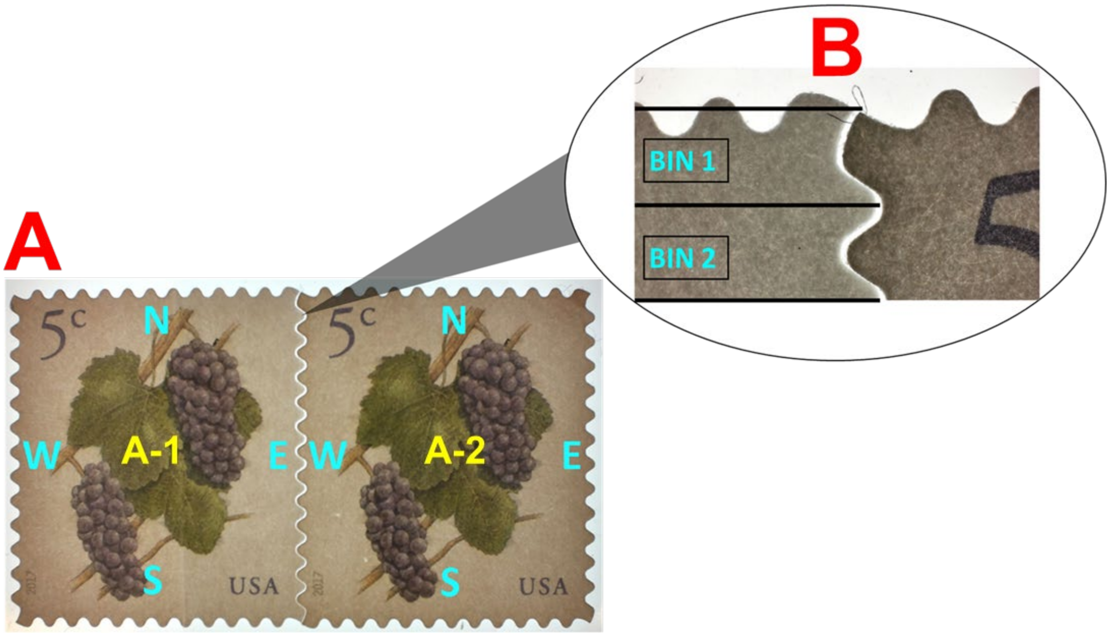


Figure 31. *(A) The alignment of a comparison pair is shown, with cardinal directions marked to indicate the comparison of the east and west edges of the pair. (B) The subdivision of the comparison edge into bins can be seen. Each bin contains an upper crest, trough, and lower crest.*

5.2.4 Data Analysis

Performance rates were used to evaluate the overall rigorousness of the method. False-positive rates, false-negative rates, sensitivity, specificity, and overall accuracy were calculated for each of the two examiners. These metrics consider the various outcomes of the comparative analysis process (true-positive (TP), true-negative (TN), false-positive (FP), false-negative (FN), and inconclusive (IN)). Performance rates are calculated using the number of times these five outcomes occur in a dataset, and the equations used for these rates can be seen in Table 1.

Boxplots and frequency distributions were used to analyze the number of microfibers assigned to both true fits and true non-fits. Box plots graphically display the spread of the data through quartiles. These plots can visualize the spread of the number of aligning microfibers assigned to true fits and true non-fits in the data set. Frequency distributions provide a similar view of the data. They are helpful in exploring potential overlap in the data if, for example, specific number of aligning microfibers were identified in both a true non-fit comparison and a true fit comparison. These plots help investigate if a threshold could be established for fit/non-fit classification based on the number of aligning microfibers identified by an examiner in the experimental datasets with ground-truth information.

5.2.5 Method training and reaching consensus terminology and comparison criteria

After both examiners completed the first 35 comparisons, the results to that point were evaluated. There was a significant discrepancy between the number of aligning microfibers that the two examiners identified. Examiner 1 identified significantly more aligning microfibers than Examiner 2. The examiners hypothesized that this was because of Examiner 1's greater experience working with physical fit comparisons and other fibrous materials in contrast to Examiner 2 (e.g., > 3 years of experience, > 1000 comparisons performed vs. < 1 year, 120 comparisons performed). Therefore, the first 35 comparisons were repeated by both examiners together to better understand and identify aligning microfibers. The dataset was also independently reviewed by a third examiner. After these consensus comparisons were completed, the team evaluated the sources of discrepancy and further defined the comparison criteria, the reporting and terminology, and methodology. Once the improved method was established, the two examiners independently completed the remaining 65 comparisons.

5.2.6 Sample Imaging

All samples were imaged as pairs using a portable microscope (MiScope® (Zarbeco; Succasunna, NJ)). Images were captured at 40X magnification for each comparison bin using both transmitted light and reflected UV light to best visualize the presence of microfibers.

5.3 RESULTS AND DISCUSSION

Unlike previous materials that our group has investigated, the lack of unique features imparted on these samples through the machine-cut separation method caused microfiber alignment to be a critical feature in determining the presence of a fit or a non-fit. Since it is assumed that the distribution of microfibers on a piece of paper material is random, it should be expected that no aligning microfibers would be observed in true non-fitting comparison pairs. Therefore, the observation of aligning fibers along the fracture edge of a comparison pair should indicate that the two samples originated from adjacent positions on the same sheet of postage stamps.

However, this assumption was found to be invalid after conducting these comparisons with multiple examiners. Microfiber alignment was identified in some true non-fitting comparison pairs by both examiners (ranging from 1 to 11 aligning microfibers). In Figure 32, the distribution of aligning microfibers observed for both true fits and true non-fits by both examiners can be observed in boxplot format. While the average number of aligning microfibers for true non-fitting stamps was greater than zero for both examiners (2), the average number of aligning microfibers in true fitting stamps (59) was still far greater than the number for true non-fits. The vast difference between the number of microfibers identified between examiners was attributed to the fact that, at the time, Examiner 1 was much more experienced in conducting physical fit examinations with other materials and had done some preliminary examinations of paper. On the other hand, Examiner 2 was much less experienced with physical fit comparisons and had no prior experience with paper materials. Examiner 1 was better suited at noting small details, such as microfibers, and therefore identified a greater number of these microfibers. When the previous review was conducted after the first 35 comparisons, it showed that although Examiner 1 was identifying the same microfibers locations as Examiner 2, there were additional alignments only found by the most experienced one.

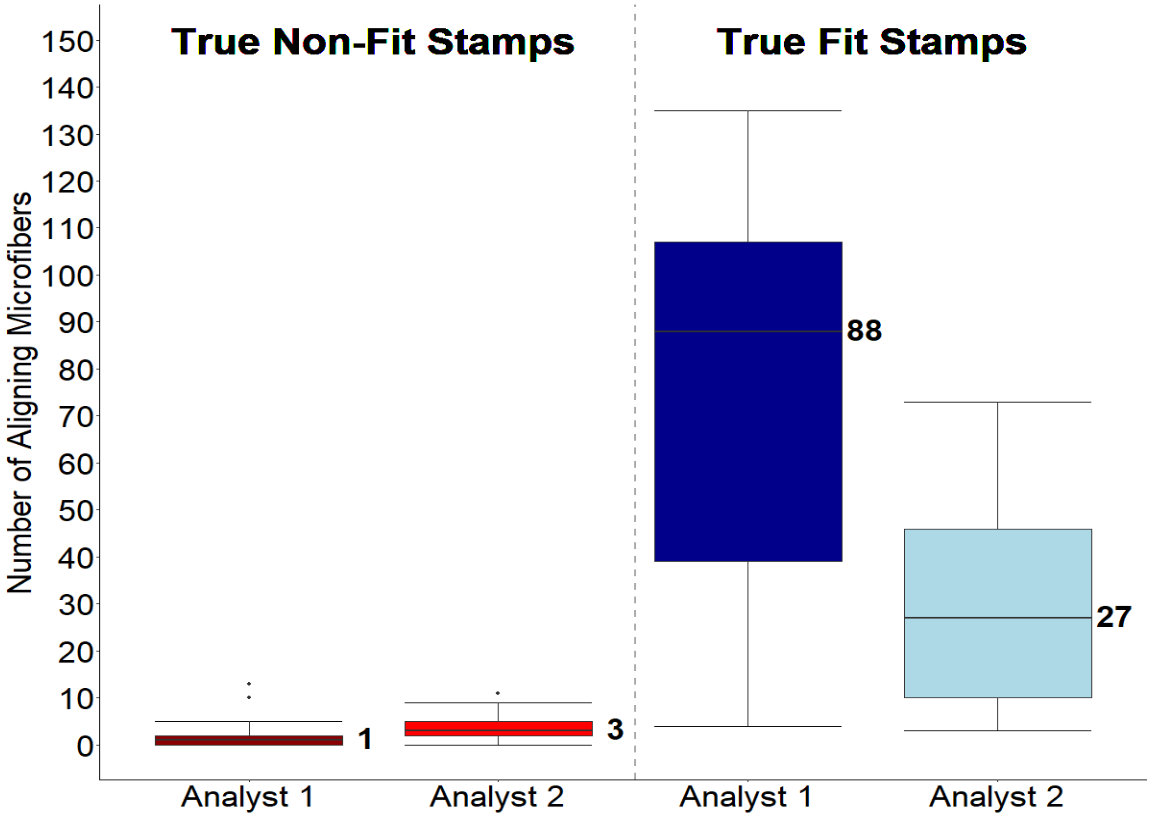


Figure 32. *Boxplots showing the distribution of the number of aligning microfibers identified by both examiners in true fits and true non-fits*

Figure 33 shows examples of aligning microfibers and misaligning microfibers in true fits and true non-fits. Figure 33A represents a typical true-fitting comparison bin. There are several instances of aligning microfibers, including a substantially large microfiber in the center of the bin. The presence of these long aligning microfibers increases an examiner's confidence in the presence of a physical fit, especially if this kind of alignment is observed across several comparison bins. Figure 33B provides an example of microfiber misalignment in a true non-fitting comparison pair. A long microfiber approaches the comparison edge of the left-hand stamp from the southwest. That fiber does not continue onto the right-hand stamp, rather abruptly terminating at the edge.

When an extremely long or thick microfiber abruptly terminates at the comparison edge, it can indicate a non-fit. Finally, Figure 33C showcases an example of microfiber alignment in a true non-fitting comparison pair. The example provided here is extreme, as three aligning microfibers were observed in a single comparison bin. This does highlight the need for the examiner to consider the entirety of the comparison edge, rather than a single bin when formulating their conclusion.

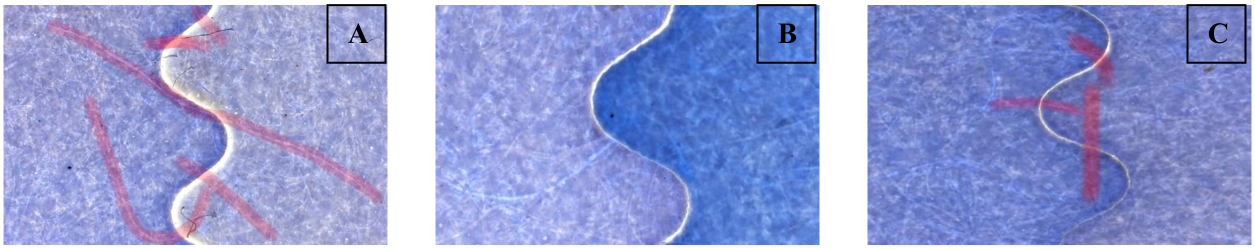


Figure 33. *UV images of a true fitting comparison bin (left, A) and two true non-fitting comparison bins (B and C). Instances of microfiber alignment are highlighted in red.*

The accuracy of this method was found to be high, indicating promise in using the number of aligning microfibers across the edge of two stamps as a metric of the quality of a physical fit. Performance rates for both examiners can be seen in Table 18. Of the ten total misclassifications made by both examiners, only one misclassification was shared between both examiners. One true fitting comparison pair was classified as a non-fit by both examiners. For this comparison, Examiner 1 observed four aligning microfibers, while Examiner 2 observed five aligning microfibers. The four aligning microfibers identified by Examiner 1 were located in the trough of Bin 4, the upper crest and trough of Bin 5, and the upper crest of Bin 11. The examiner noted the distinct misalignment of microfibers in Bins 1, 2, and 6 in his notes. The five aligning microfibers documented by Examiner 2 were located in the crest of Bin 1, the troughs of Bins 3, 5, and 8, and the crest of Bin 13.

Table 18. *Performance rate table for both examiners, including accuracy, sensitivity, and specificity, using bin-by-bin microfiber alignment as the primary decision-making factor.*

	Examiner 1 Reported Fit	Examiner 1 Reported Non-Fit	Examiner 1 Reported Inconclusive	Examiner 2 Reported Fit	Examiner 2 Reported Non-Fit	Examiner 2 Reported Inconclusive	Total Comparisons
True Fit	48 (98% True Positive)	1 (2% False Negative)	0 (0% Inconclusive)	45 (92% True Positive)	4 (8% False Negative)	0 (0% Inconclusive)	49
True Non-Fit	2 (4% False Positive)	48 (94% True Negative)	1 (2% Inconclusive)	1 (2% False Positive)	49 (96% True Negative)	1 (2% Inconclusive)	51
Accuracy		96%			94%		

For the other three false negative misclassifications made by Examiner 2, the examiner observed six, four, and three aligning microfibers across the comparison edges of each pair. Examiner 2 observed 11 aligning microfibers across the comparison edge for the single false positive misclassification. Examiner 1 observed 13 and 10 aligning microfibers on the surface of the two false positive misclassifications. It should be noted that these two comparisons were the second and third total comparisons performed in this dataset, therefore, the examiner was not yet accustomed to the typical number of aligning microfibers found in a true-fitting pair and was still assuming that no aligning microfibers would be seen in true non-fitting pairs. Additionally, these two comparisons were assigned the largest numbers of aligning microfibers for the true non-fit comparison pairs by Examiner 1. No other true non-fitting comparison pair exhibited more than five aligning microfibers.

When analyzing the number of aligning microfibers identified in both true fits and true non-fits by both examiners, it was noted that while aligning microfibers were identified in true non-fits, they were present in far fewer amounts compared to true fits. Despite this disproving the common assumption that no aligning microfibers would be identified in true non-fits, it presents an intriguing possibility that a threshold could be established to determine the fit/non-fit classification for an unknown comparison pair based on the number of aligning microfibers identified by the examiner. To investigate possible choices for a threshold based on the data collected in this study, the number of aligning microfibers for true fits and true non-fits for both examiners were plotted in a frequency distribution graph, which can be seen in Figure 34.

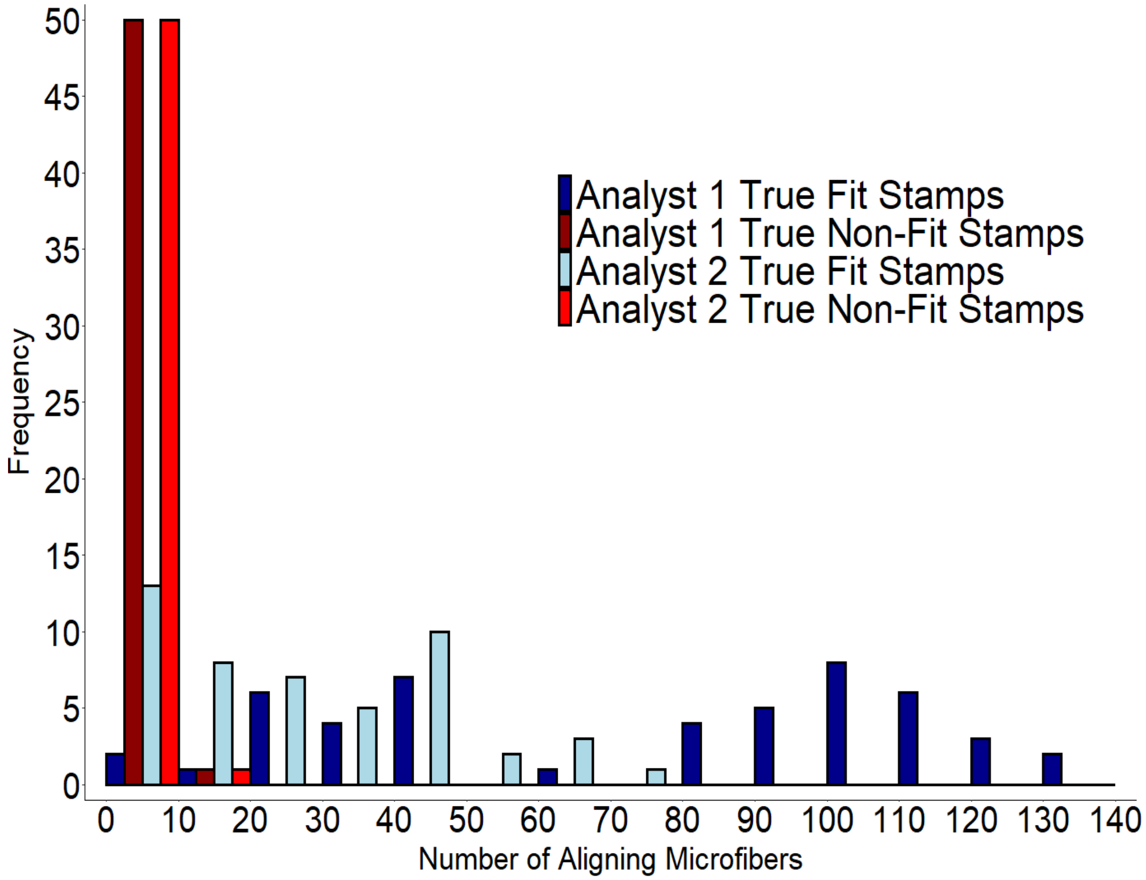


Figure 34. *Frequency distribution for the number of aligning microfibers for true fits (blue) and true non-fits (red) for Examiner 1 (dark) and Examiner 2 (light).*

When discussing the application of a threshold, the examiners noted that they subconsciously had used a mental threshold of ten aligning microfibers as an indicator of the presence of a physical fit. This was interesting because they had come up with that criteria independently with no external input. This was also supported by the frequency distribution seen above, as the vast majority of true non-fitting comparison pairs were observed to contain fewer than ten aligning microfibers. Therefore, a threshold of ten aligning microfibers was applied to the previously collected data. The threshold completely replaced the final conclusion reported by each examiner to evaluate how the sole metric of microfiber alignment performs. If ten or more aligning microfibers were observed in a given comparison, that comparison was reported as a fit. If the examiner observed less than ten aligning microfibers, that comparison was reported as a non-fit. After this threshold was applied, the performance rates for each examiner were recalculated, which can be seen in Table 19.

Table 19. *Performance rates for both examiners when using a threshold of ten aligning microfibers for classification*

	Examiner 1 Reported Fit	Examiner 1 Reported Non-Fit	Examiner 1 Reported Inconclusive	Examiner 2 Reported Fit	Examiner 2 Reported Non-Fit	Examiner 2 Reported Inconclusive	Total Comparisons
True Fit	48 (98% True Positive)	1 (2% False Negative)	0 (0% Inconclusive)	41 (84% True Positive)	8 (16% False Negative)	0 (0% Inconclusive)	49
True Non-Fit	2 (4% False Positive)	49 (96% True Negative)	0 (0% Inconclusive)	1 (2% False Positive)	50 (98% True Negative)	0 (0% Inconclusive)	51
Accuracy		97%			91%		

Using this threshold, the accuracy for Examiner 1 improves slightly, as the true non-fit that was classified as inconclusive by the examiner is now classified as a non-fit by the threshold. However, accuracy decreased for Examiner 2. While one inconclusive misclassification was corrected and deemed a non-fit by the threshold, the false negative rate for Examiner 2 doubles. This is due to four true-fitting comparisons within the first 35 comparisons conducted before the consensus training that were correctly classified as a fit by Examiner 2 but were not reported to display ten or more aligning microfibers. Examiner 2 observed four, nine, seven, and nine aligning microfibers for these comparisons.

To statistically evaluate the best possible threshold for this dataset, performance rates were calculated for individual thresholds. The false positive and false negative rates were plotted against each other, which can be seen in Figure 35 for Examiner 1 and Figure 36 for Examiner 2. The most interesting area of these plots is the point or series of points where the two rates intersect. This point represents the threshold(s) with the lowest combination of false negative and false positive rates and, therefore, the best performance. The plots indicate that the optimal thresholds are in the range of 6-13 aligning microfibers for Examiner 1 and 6-7 aligning microfibers for Examiner 2. A range is provided because the performance is comparable for every threshold in that range. This data corroborates the selection of 10 aligning microfibers as an acceptable threshold, but it is not the most optimal for this dataset.

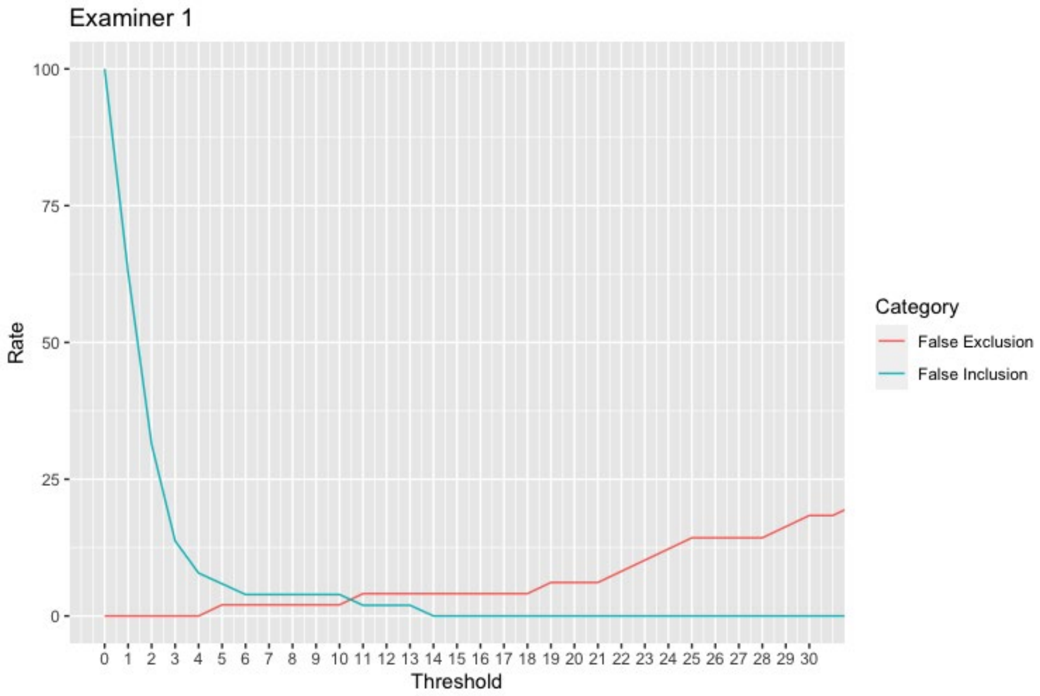


Figure 35. *Plot of false positive rate and false negative rate for Examiner 1 versus microfiber alignment thresholds from 1 to 30*

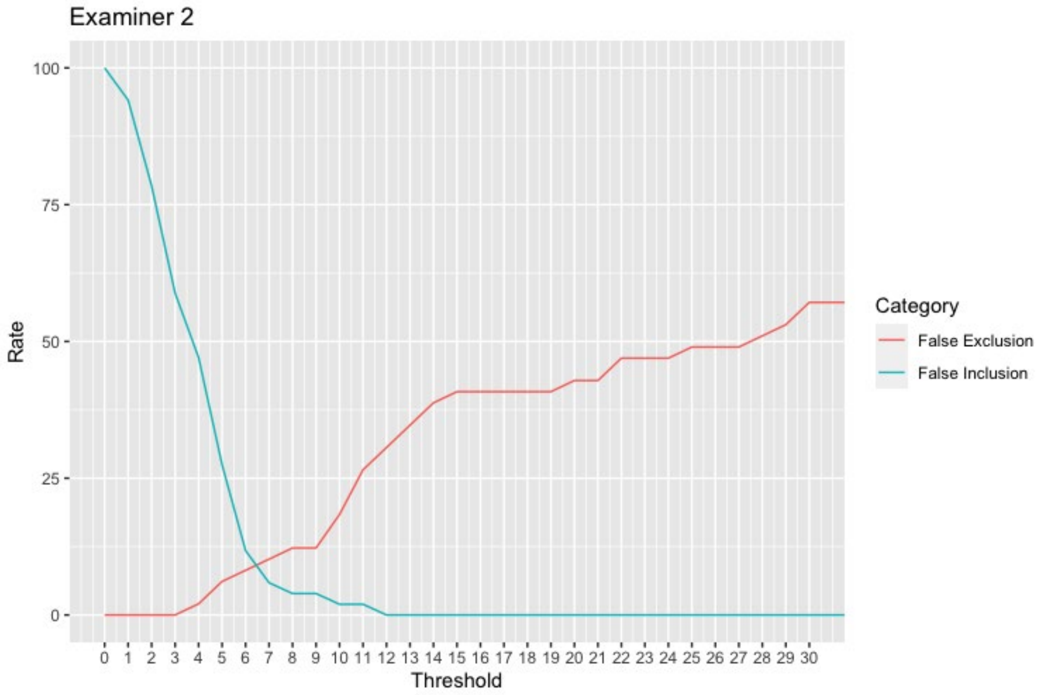


Figure 36. *Plot of false positive rate and false negative rate for Examiner 2 versus microfiber alignment thresholds from 1 to 30*

Receiver-Operator Characteristic (ROC) curves were made to further investigate the performance of this method for the dataset. The ROC curves for Examiner 1 and Examiner 2 can be seen in Figure 37. Examiner 1 has an AUC of 0.97, while Examiner 2 has an AUC of 0.92. Both of these values indicate very good performance by the method at distinguishing between fits and non-fits. It must be restated that this is a relatively small dataset, so future work should be done to expand the sample size and gain an improved understanding of microfiber alignment in postage stamps.

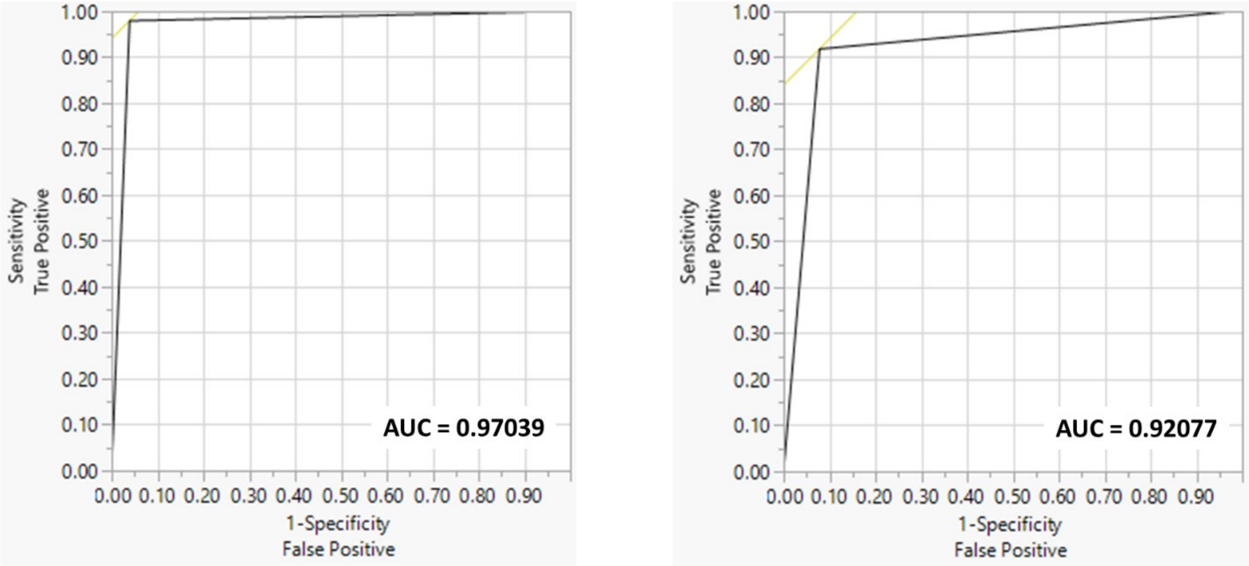


Figure 37. *ROC curves for Examiner 1 (left) and Examiner 2 (right)*

To explore the effect of training on the dataset, the results of the first 35 comparisons for both examiners were replaced with the results of the consensus training previously discussed. The updated performance rates reflecting the consensus training and using the remaining 65 examiners' reported conclusions can be seen in Table 20. These results show that the training improved examiner performance significantly. Accuracy improved by 4% for both examiners. Only two total misclassifications were reported, both for Examiner 2. This examiner misclassified two true non-fitting comparison pairs: one as a false positive and the other as an inconclusive. The examiner observed eleven and nine aligning microfibers, respectively, for these pairs. The effect of applying a ten-aligning microfiber threshold to this data can be seen in Table 21. Notably, this does not affect the performance of Examiner 1, though it does marginally improve the performance of Examiner 2 by correcting the inconclusive misclassification. However, the threshold did not correct the false positive misclassification as the examiner observed greater than ten aligning microfibers for this comparison pair.

Table 20. *Performance rates for both examiners using consensus training data for the first 35 comparisons*

	Examiner 1 Reported Fit	Examiner 1 Reported Non-Fit	Examiner 1 Reported Inconclusive	Examiner 2 Reported Fit	Examiner 2 Reported Non-Fit	Examiner 2 Reported Inconclusive	Total Comparisons
True Fit	49 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	49 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	49
True Non-Fit	0 (0% False Positive)	51 (100% True Negative)	0 (0% Inconclusive)	1 (2% False Positive)	49 (96% True Negative)	1 (2% Inconclusive)	51
Accuracy	100%			98%			

Table 21. *Performance rates for both examiners using consensus training data for the first 35 comparisons and a threshold of ten aligning microfibers for classification*

	Examiner 1 Reported Fit	Examiner 1 Reported Non-Fit	Examiner 1 Reported Inconclusive	Examiner 2 Reported Fit	Examiner 2 Reported Non-Fit	Examiner 2 Reported Inconclusive	Total Comparisons
True Fit	49 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	49 (100% True Positive)	0 (0% False Negative)	0 (0% Inconclusive)	49
True Non-Fit	0 (0% False Positive)	51 (100% True Negative)	0 (0% Inconclusive)	1 (2% False Positive)	50 (98% True Negative)	0 (0% Inconclusive)	51
Accuracy	100%			99%			

Again, to statistically evaluate the best possible threshold for this dataset, performance rates were calculated for individual thresholds, and the false positive and false negative rates were plotted against each other, which can be seen in Figure 38 for Examiner 1 and Figure 39 for Examiner 2. The plots indicate that the new optimal thresholds are 8-13 aligning microfibers for Examiner 1 and 12-13 aligning microfibers for Examiner 2. This data corroborates the original selection of 10 aligning microfibers as an acceptable threshold. Still, it is not optimal for this dataset, as a selection of 12, for example, would provide the best possible results.

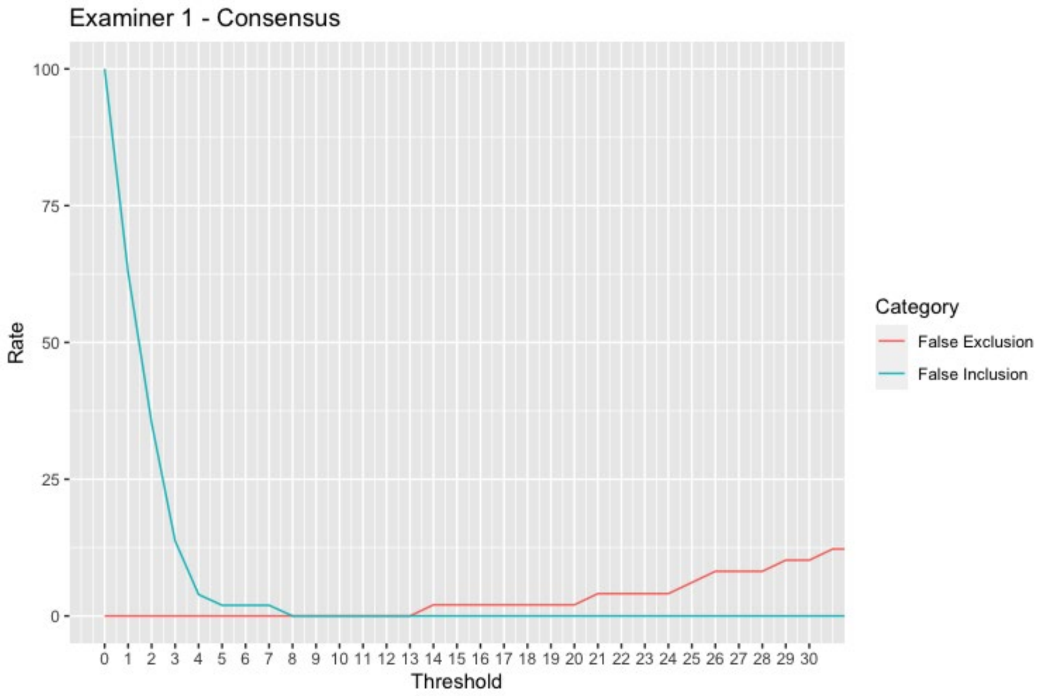


Figure 38. *Plot of false positive rate and false negative rate for Examiner 1 versus microfiber alignment thresholds from 1 to 30, using consensus training data for the first 35 comparisons*

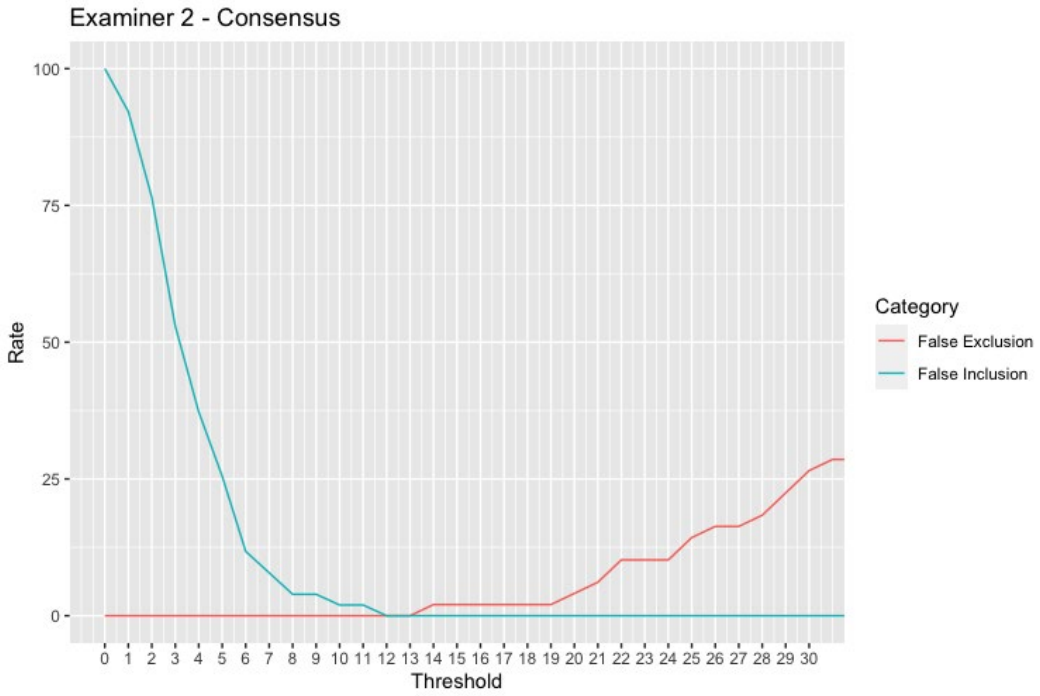


Figure 39. *Plot of false positive rate and false negative rate for Examiner 2 versus microfiber alignment thresholds from 1 to 30, using consensus training data for the first 35 comparisons*

ROC curves for the consensus training data can be seen in Figure 40 for Examiner 1 and Examiner 2. For both examiners, AUC improved when the consensus training was considered. The AUC for Examiner 1 improved from 0.97 to 1.0, while Examiner 2's AUC improved from 0.92 to 0.99. This further cements the importance of training and experience for physical fit comparisons and microfiber alignment identification. These AUC values also indicate that the use of microfiber alignment as a metric for quantifying the quality of a physical fit is excellent at differentiating between fits and non-fits for this dataset.

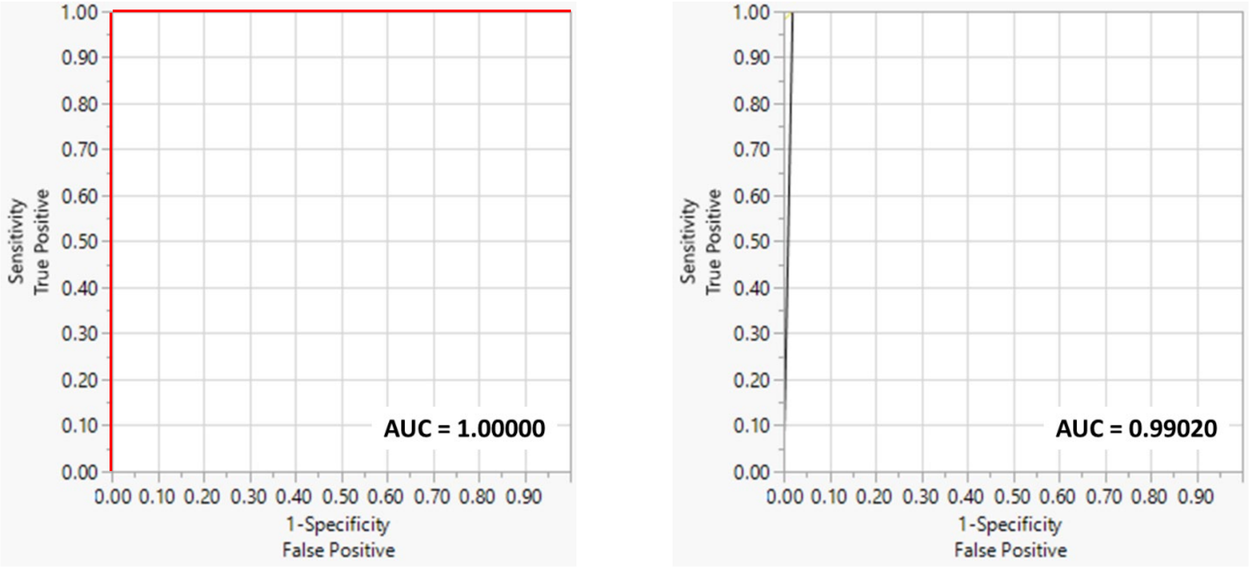


Figure 40. *ROC curves for Examiner 1 (left) and Examiner 2 (right) using the consensus training data for the first 35 comparisons*

5.4 CONCLUSIONS

A systematic, quantitative approach for assessing the quality of a physical fit for postage stamps was evaluated in this study. A set of 100 comparison pairs of stamp edges composed of true fits and true non-fits was used to quantify the occurrence of micro-fiber alignments (MFA) and document their features. The ground truth of the dataset was kept blind from the two examiners so the error rates could be assessed. This study also served to test the assumption that no aligning microfibers would be observed in a true non-fitting pair of stamps due to the random assortment of microfibers on the surface of paper materials. Unlike previously studied materials, no distinctive macroscopic edge alignment was observed among all comparisons, regardless of the ground truth, illustrating the need to use microfiber alignment to make inferences regarding the presence of a fit.

The assumption that microfiber alignment occurs only on the edges of true fitting pairs was disproved in this study. These findings agreed with observations made for paper specimens described in the previous chapter. True non-fitting stamp edges were found to exhibit some microfiber alignment. However, it was observed that the number of aligning microfibers found on the edges of true fitting pairs greatly exceeded the number of aligning microfibers found on the edges of true non-fitting pairs. These results indicate that a threshold of aligning microfibers could be established to inform the examiner's opinion and provide a more objective means of classifying an unknown comparison pair. Nonetheless, a larger dataset with various stamp types is required to generalize these findings.

The method produced high accuracies, ranging between 91% and 100% for both examiners. While inter-examiner variance was observed for the number of aligning microfibers documented by each examiner, especially for true fitting comparison pairs, this variation had minimal effect on the overall accuracy.

This study highlights the importance of training and experience on examiner performance. The examiner with more experience conducting physical fit comparisons of other materials, such as duct tape and textiles, was able to identify a greater number of aligning microfibers than the examiner with less experience in physical fit comparisons. Once the examiners underwent a joint consensus training involving the re-evaluation of several comparison pairs, the performance of the second examiner significantly improved. They also began identifying significantly more aligning microfibers per true fit than they did before the training. This training was critical in mitigating inter-examiner variation in microfiber identification. This training event and the documentation of microfiber location in each comparison bin provide the framework for a standardized, transparent peer review process that can further enhance the rigorousness of this method.

While this method serves as a basis to establish data-driven qualitative and quantitative descriptors of the quality of a physical fit for postage stamps, more extensive studies need to be performed in the future to corroborate and expand the findings of the study. Future work should explore the application of thresholds for classification purposes. By incorporating larger datasets, threshold selection should become more apparent, as would the inter-examiner variance for microfiber identification, which is vital for assessing the selection of a classification threshold and the method in its entirety.

6. OVERALL CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSIONS FOR THE PRESENT STUDY

This project sought to assess the validity and reliability of physical fit comparisons for three materials: textiles, paper, and postage stamps. This was accomplished, in part, by the development of consensus-based, systematic, and quantitative methodologies for each material. This included the standardization and definition of relevant features that can influence decision-making during physical fit comparisons. These metrics all address gaps in the discipline as outlined by the National Institute of Justice and the National Institute of Standards and Technology. The materials used in this study are commonly found in criminal investigations and are frequently submitted to forensic laboratories for physical fit analysis. Despite this, there is a dearth of literature regarding physical fit comparisons for these materials, and potential error rates are not fully understood.

The development of a quantitative, systematic approach for assessing physical fits of textiles was successful. The results first highlighted the need to understand the suitability of different fabric compositions to undergo physical fit analysis. Some compositions of fabric simply deform too extremely to produce accurate results. The proposed method produced high overall accuracy, ranging from 88% to 100% for different types of fabrics once suitability issues were implemented. Standardized terminology and relevant features were defined, which serves as a basis for a more uniform reporting and decision-making process. Training and experience were also important. After almost 200 comparisons were completed, examiner performance improved, as did intra-examiner accuracy on re-analyzed samples.

The development of a systematic method for assessing physical fits of paper was also successful. While aligning microfibers were observed in true non-fitting comparisons, the overlap between the number of aligning microfibers in true fitting pairs and true non-fitting pairs was minimal. When this factor was combined with relevant features observed along the comparison edges, such as letter alignment and edge shape alignment, fits and not-fits determinations showed minimal misclassifications (92.5% - 100% accuracy). Additionally, the implementation of a classification threshold improved the objectivity of the method while slightly improving accuracy in the process. No substantial differences were observed regarding paper brightness and manufacturer or printer model.

Finally, microfiber alignment was also observed on the edges of true non-fitting postage stamps. This proved challenging because of the lack of other distinguishing features along the comparison edge that could be used to classify a pair as either a fit or a non-fit. Despite this, there was a significant separation between the number of aligning microfibers observed in true fits and true non-fits. Due to this difference, fits and non-fits could still be classified accurately while minimizing misclassifications. This study highlighted the need for the definition of terminology and feature identification, as examiner accuracy improved once a consensus review was done to better define microfiber alignment. Additionally, the wide separation between the number of microfibers identified in true fits and true non-fits proved promising for the implementation of the classification threshold for classifying fits and non-fits. A threshold of 10-12 aligning microfibers for classifying an unknown pair as a fit improved accuracy and provided a more objective means of classification.

Overall, these methods for comparative analysis of fractured materials provide an empirical analysis of physical fit comparisons for textiles, paper, and postage stamps that were sorely lacking in the literature. Performance rates for physical fit comparisons were discussed, and these methods display high accuracy for separating fits and non-fits. This research defined standardized terminology for these materials and defined systematic approaches for analyzing them, which increase objectivity and provide a means of straightforward peer review that is conducive to the current needs of practitioners.

6.2 RECOMMENDATIONS FOR FUTURE WORK

The findings of this study also illustrated areas that could be addressed by future research. These areas include:

a. The establishment of an inter-laboratory study for textile physical fits so that active practitioner feedback can be incorporated into the method. This would also aid in the implementation of the method into forensic laboratories.

b. The expansion of the textile physical fit dataset to incorporate simulated casework samples to test the effects of blood, dirt, mud, bleach, laundering, and more on the quality of a physical fit.

c. The extension of sample size for the paper physical fit set to increase the number of comparison pairs for different brands and types of paper and manufacturers and models of printers. Additional paper types that should be studied include notebook paper, legal paper, and construction paper.

d. The expansion of the postage stamp dataset to include different types of postage stamps, including different qualities of stamps and stamps from various sources, such as coils or vending machines.

e. Probabilistic interpretation of the data to aid examiners assess the weight of the evidence.

7. REFERENCES

1. Schwartz, T., Rothenberg, D. & Clark, B. Trace Evidence Recognition, Collection, and Preservation. in *Handbook of Trace Evidence Analysis* (eds. Desiderio, V., Taylor, C. & Nic Daeid, N.) 1–31 (Wiley, 2021).

2. National Academy of Sciences. *Strengthening forensic science in the United States: A path forward*. *Strengthening Forensic Science in the United States: A Path Forward* (The National Academies, 2009). doi:10.17226/12589.

3. President's Council of Advisors on Science and Technology. Report to the President - Forensic Science in Criminal Courts: Ensuring Scientific Validity. *Exec. Off. Pres. Pres. Counc. Advis. Sci. Technol.* **1**, 1–160 (2016).

4. Organization of Scientific Area Committees. Development of Quantitative Assessment and Evaluation of Error Rates in Physical Fit Determinations of Trace Materials. *Natl. Inst. Sci. Technol.* (2018).

5. National Institute of Justice. Forensic Technology Working Group Operational Requirements, November 2018. (2019).

6. Gross, S. *2020 Physical Fits Survey*. (2020).

7. Prusinowski, M., Brooks, E. & Trejos, T. Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits. *Forensic Sci. Int.* **307**, 110103 (2020).

8. Trejos, T., Koch, S. & Mehltretter, A. Scientific foundations and current state of trace evidence—A review. *Forensic Chem.* **18**, 100223 (2020).

9. Van Amber, R. Apparel and Household Textiles and Their Role in Forensics. in *Forensic Textile Science* 15–26 (Elsevier, 2017).

10. Dann, T. & Malbon, C. Tearing or Ripping of Fabrics. *Forensic Text. Sci.* 169–180 (2017) doi:10.1016/B978-0-08-101872-9.00008-X.

11. Koch, S. & Nehse, K. Fibers. in *Handbook of Trace Evidence Analysis* (eds. Desiderio, V., Taylor, C. & Nic Daeid, N.) 322–376 (Wiley, 2021).

12. Saferstein, R. Document Examination. in *Forensic Science: From the Crime Scene to the Crime Lab* 439–457 (Pearson, 2016).

13. Ramotowski, R. The History, Manufacturing, and Analysis of Paper. (2002).

14. Biermann, C. J. Pulping Fundamentals. in *Handbook of Pulping and Papermaking* 55–71 (Elsevier, 1996).

15. Bajpai, P. Green chemistry and sustainability in pulp and paper industry. *Green Chem. Sustain. Pulp Pap. Ind.* 1–258 (2015) doi:10.1007/978-3-319-18744-0.

16. McKinstry, E. A. Fracture match - a case study. *AFTE J.* **30**, 343–344 (1998).

17. Klein, A., Nedivi, L. & Silverwater, H. Phvsical match of fragmented bullets. *J. Forensic Sci.* **45**, 722–727 (2000).

18. Robinson, M. Comparison of gunstock parts to barreled action. *Herpetol. Rev.* 65–69 (1976).

19. Dawood, B. *et al.* Quantitative matching of forensic evidence fragments utilizing 3D microscopy analysis of fracture surface replicas. *J. Forensic Sci.* **67**, 899–910 (2022).

20. Vanderkolk, J. R. Identifying Consecutively Made Garbage Bags Through Manufactured Characteristics. *J. Forensic Identif.* **45**, 38–50 (1995).

21. Von Bemen, U. G. & Blunt, L. K. R. Physical comparison of plastic garbage bags and sandwich bags. *J. Forensic Sci.* **28**, 644–654 (1983).

22. VanHoven, H. A. & Fraysier, H. D. The matching of automotive paint chips by surface striation alignment. *J Forensic Sci* **28**, 463–467 (1983).

23. Walsh, K. & Gordon, A. Pattern matching of a paint flake to its source. *AFTE J.* **33**, 143–145 (2001).

24. Osterburg, J. W. *The Crime Laboratory: Case Studies of Scientific Criminal Investigation*. (Indiana University Press, 1968).

25. Bradley, M. J. *et al.* A validation study for duct tape end matches. *J. Forensic Sci.* **51**, 504–508 (2006).

26. Bradley, M. J., Gauntt, J. M., Mehltretter, A. H., Lowe, P. C. & Wright, D. M. A validation study for vinyl electrical tape end matches. *J. Forensic Sci.* **56**, 606–611 (2011).

27. Tulleners, F. & Braun, J. The Statistical Evaluation of Torn and Cut Duct Tape Physical End Matching. (2011).

28. Mccabe, K. R., Tulleners, F. A., Braun, J. V., Currie, G. & Gorecho, E. N. A Quantitative Analysis of Torn and Cut Duct Tape Physical End Matching. *J. Forensic Sci.* **58**, 34–42 (2013).

29. van Dijk, C. D., van Someren, A., Visser, R. & Sjerps, M. Evidential value of duct tape comparison using loopbreaking patterns. *Forensic Sci. Int.* **332**, 111178 (2022).

30. Kemp, S. E., Carr, D. J., Kieser, J., Niven, B. E. & Taylor, M. C. Forensic evidence in apparel fabrics due to stab events. *Forensic Sci. Int.* **191**, 86–96 (2009).

31. Sloan, K., Fergusson, M. & Robertson, J. Textile damage examinations on the cutting edge–an Australian perspective. *Aust. J. Forensic Sci.* **50**, 682–688 (2018).

32. Cowper, E. J., Carr, D. J., Horsfall, I. & Fergusson, S. M. The effect of fabric and stabbing variables on severance appearance. *Forensic Sci. Int.* **249**, 214–224 (2015).

33. Ukovich, A. & Ramponi, G. Features for the reconstruction of shredded notebook paper. *Proc. - Int. Conf. Image Process. ICIP* **3**, 93–96 (2005).

34. Kleber, F., Diem, M. & Sablatnig, R. Torn Document Analysis as a Prerequisite for Reconstruction. *VSMM 2009 - Proc. 15th Int. Conf. Virtual Syst. Multimed.* 143–148

(2009) doi:10.1109/VSMM.2009.27.

35. Lotus, R., Varghese, J. & Saudia, S. An approach to automatic reconstruction of apictorial hand torn paper document. *Int. Arab J. Inf. Technol.* **13**, 457–461 (2016).

36. Justino, E., Oliveira, L. S. & Freitas, C. Reconstructing shredded documents through feature matching. *Forensic Sci. Int.* **160**, 140–147 (2006).

37. De Smet, P. Reconstruction of ripped-up documents using fragment stack analysis procedures. *Forensic Sci. Int.* **176**, 124–136 (2008).

38. Diem, M., Kleber, F. & Sablatnig, R. Document analysis applied to fragments. 393–400 (2010) doi:10.1145/1815330.1815381.

39. Li, P., Fang, X., Pan, L., Piao, Y. & Jiao, M. Reconstruction of shredded paper documents by feature matching. *Math. Probl. Eng.* **2014**, (2014).

40. Aguilar, M. Physical Match: Uniqueness of Torn Paper. *Themis Res. J. Justice Stud. Forensic Sci.* **7**, 4 (2019).

41. Prusinowski, M. N. Assessing the reliability of physical end matching and chemical comparison of pressure sensitive tapes. (2019).

42. Brooks, E. K. Statistical Assessment of the Significance of Fracture Fits in Trace Evidence. *West Virginia Univ. Thesis* (2020).

43. Prusinowski, M., Andrews, Z., Nguyen, E. & Trejos, T. Development of Systematic Approaches for Physical Fit Comparisons of Trace Materials. in *73rd Annual AAFS Scientific Meeting* (2021).

44. Evett, I. W. A Quantitative Theory for Interpreting Transfer Evidence in Criminal Cases. *J. R. Stat. Soc.* **33**, 25–32 (1984).

45. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).

46. Brooks, E., Prusinowski, M., Gross, S. & Trejos, T. Forensic physical fits in the trace evidence discipline: A review. *Forensic Sci. Int.* **313**, 110349 (2020).

47. Nelson, D. F. Illustrating the Fit of Glass Fragments. *J. Crim. Law. Criminol. Police Sci.* **50**, 312 (1959).

48. Funk, H. J. Comparison of paper matches. *J. Forensic Sci.* **13**, 137–143 (1967).

49. Bisbing, R. E., Willmer, J. H., LaVoy, T. A. & Berglund, J. S. A fingernail identification. *AFTE J.* **12**, 27–28 (1980).

50. Dror, I. E. Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias. *Anal. Chem.* **92**, 7998–8004 (2020).

51. Quigley-McBride, A., Dror, I. E., Roy, T., Garrett, B. L. & Kukucka, J. A practical tool for information management in forensic decisions: Using Linear Sequential Unmasking-Expanded (LSU-E) in casework. *Forensic Sci. Int. Synerg.* **4**, 100216 (2022).

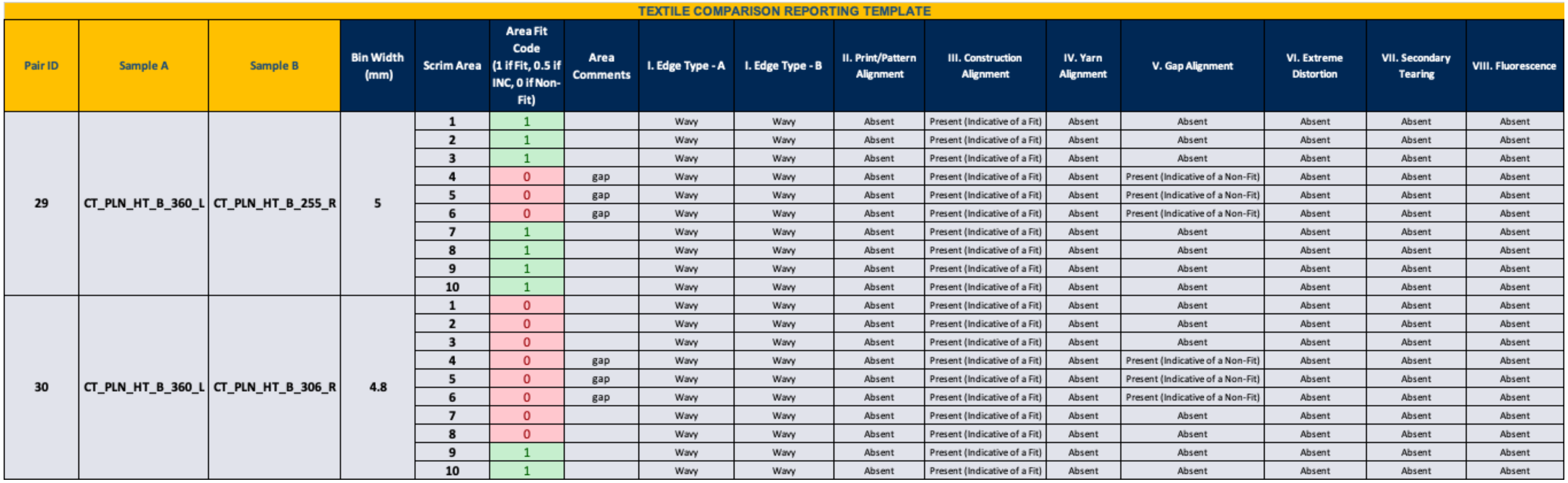
52. Tolles, J. & Meurer, W. J. Logistic regression: Relating patient characteristics to

outcomes. *JAMA - J. Am. Med. Assoc.* **316**, 533–534 (2016).

53. Greenbaum, P. J. Brighter, whiter freesheet trend offers opportunity with tradeoffs. *Pulp Pap.* **80**, (2006).

APPENDIX I. CHAPTER THREE SUPPLEMENTAL MATERIAL

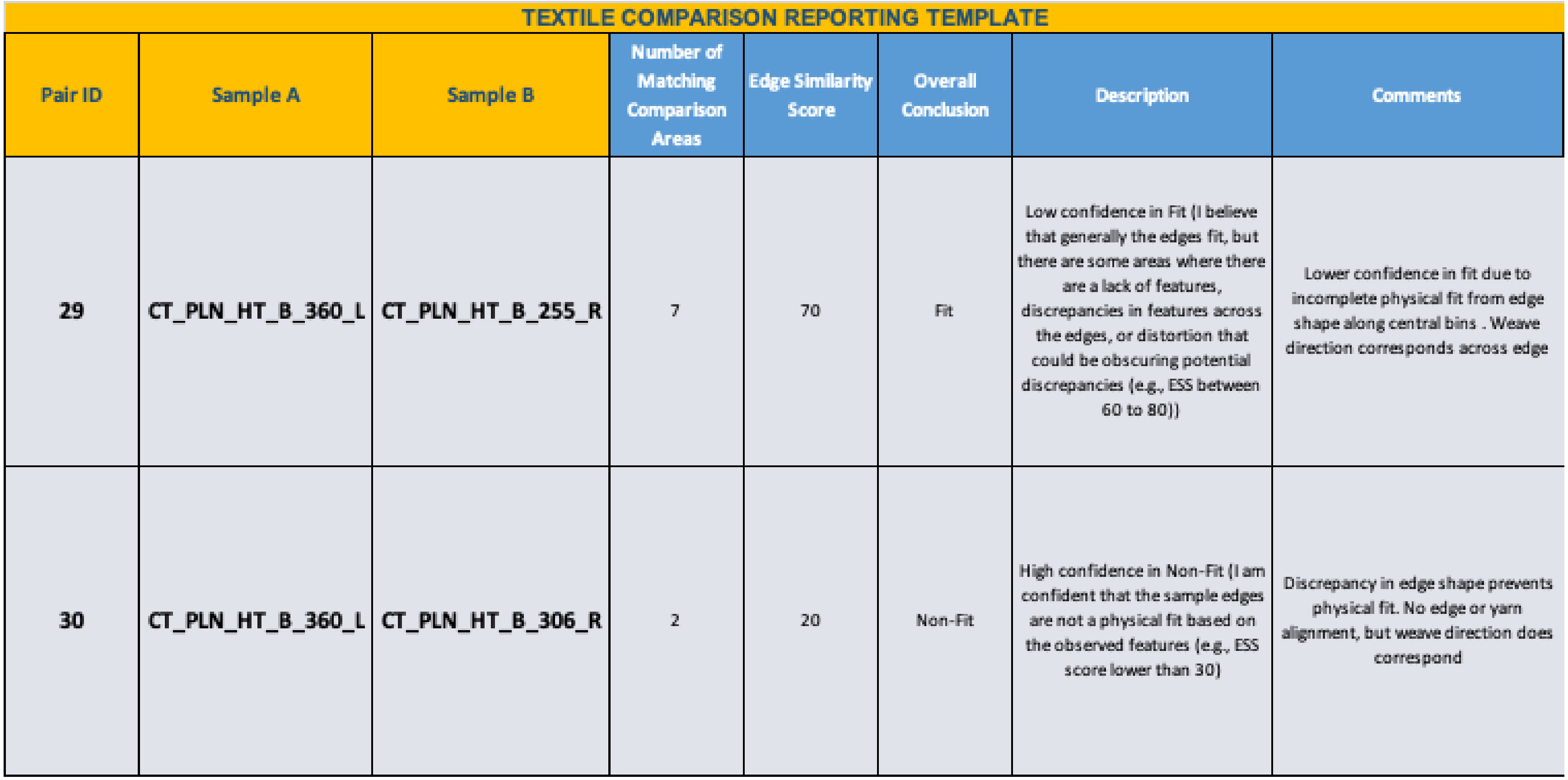
Textile Physical Fit Documentation Template Example (Part 1)



TEXTILE COMPARISON REPORTING TEMPLATE

Pair ID	Sample A	Sample B	Bin Width (mm)	Scrim Area	Area Fit Code (1 if Fit, 0.5 if INC, 0 if Non-Fit)	Area Comments	I. Edge Type - A	I. Edge Type - B	II. Print/Pattern Alignment	III. Construction Alignment	IV. Yarn Alignment	V. Gap Alignment	VI. Extreme Distortion	VII. Secondary Tearing	VIII. Fluorescence
29	CT_PLN_HT_B_360_L	CT_PLN_HT_B_255_R	5	1	1		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				2	1		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				3	1		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				4	0	gap	Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Present (Indicative of a Non-Fit)	Absent	Absent	Absent
				5	0	gap	Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Present (Indicative of a Non-Fit)	Absent	Absent	Absent
				6	0	gap	Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Present (Indicative of a Non-Fit)	Absent	Absent	Absent
				7	1		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				8	1		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				9	1		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				10	1		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
30	CT_PLN_HT_B_360_L	CT_PLN_HT_B_306_R	4.8	1	0		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				2	0		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				3	0		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				4	0	gap	Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Present (Indicative of a Non-Fit)	Absent	Absent	Absent
				5	0	gap	Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Present (Indicative of a Non-Fit)	Absent	Absent	Absent
				6	0	gap	Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Present (Indicative of a Non-Fit)	Absent	Absent	Absent
				7	0		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				8	0		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				9	1		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent
				10	1		Wavy	Wavy	Absent	Present (Indicative of a Fit)	Absent	Absent	Absent	Absent	Absent

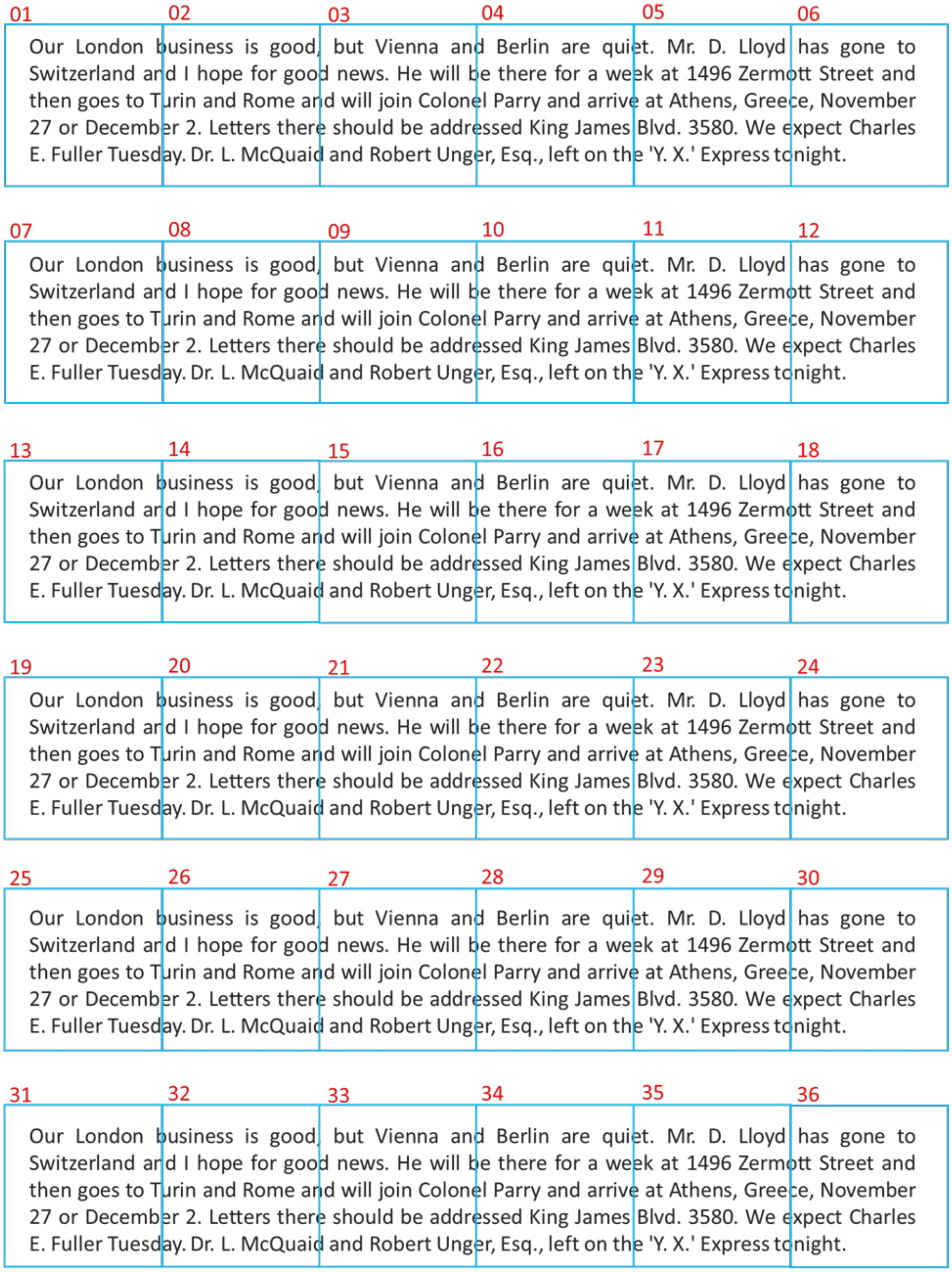
Textile Physical Fit Documentation Template Example (Part 2)



TEXTILE COMPARISON REPORTING TEMPLATE

Pair ID	Sample A	Sample B	Number of Matching Comparison Areas	Edge Similarity Score	Overall Conclusion	Description	Comments
29	**CT_PLN_HT_B_360_L**	**CT_PLN_HT_B_255_R**	7	70	Fit	Low confidence in Fit (I believe that generally the edges fit, but there are some areas where there are a lack of features, discrepancies in features across the edges, or distortion that could be obscuring potential discrepancies (e.g., ESS between 60 to 80))	Lower confidence in fit due to incomplete physical fit from edge shape along central bins . Weave direction corresponds across edge
30	**CT_PLN_HT_B_360_L**	**CT_PLN_HT_B_306_R**	2	20	Non-Fit	High confidence in Non-Fit (I am confident that the sample edges are not a physical fit based on the observed features (e.g., ESS score lower than 30)	Discrepancy in edge shape prevents physical fit. No edge or yarn alignment, but weave direction does correspond

APPENDIX II. CHAPTER FOUR SUPPLEMENTAL MATERIAL



Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott Street and then goes to Turin and Rome and will join Colonel Parry and arrive at Athens, Greece, November 27 or December 2. Letters there should be addressed King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the 'Y. X.' Express tonight.

Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott Street and then goes to Turin and Rome and will join Colonel Parry and arrive at Athens, Greece, November 27 or December 2. Letters there should be addressed King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the 'Y. X.' Express tonight.

Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott Street and then goes to Turin and Rome and will join Colonel Parry and arrive at Athens, Greece, November 27 or December 2. Letters there should be addressed King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the 'Y. X.' Express tonight.

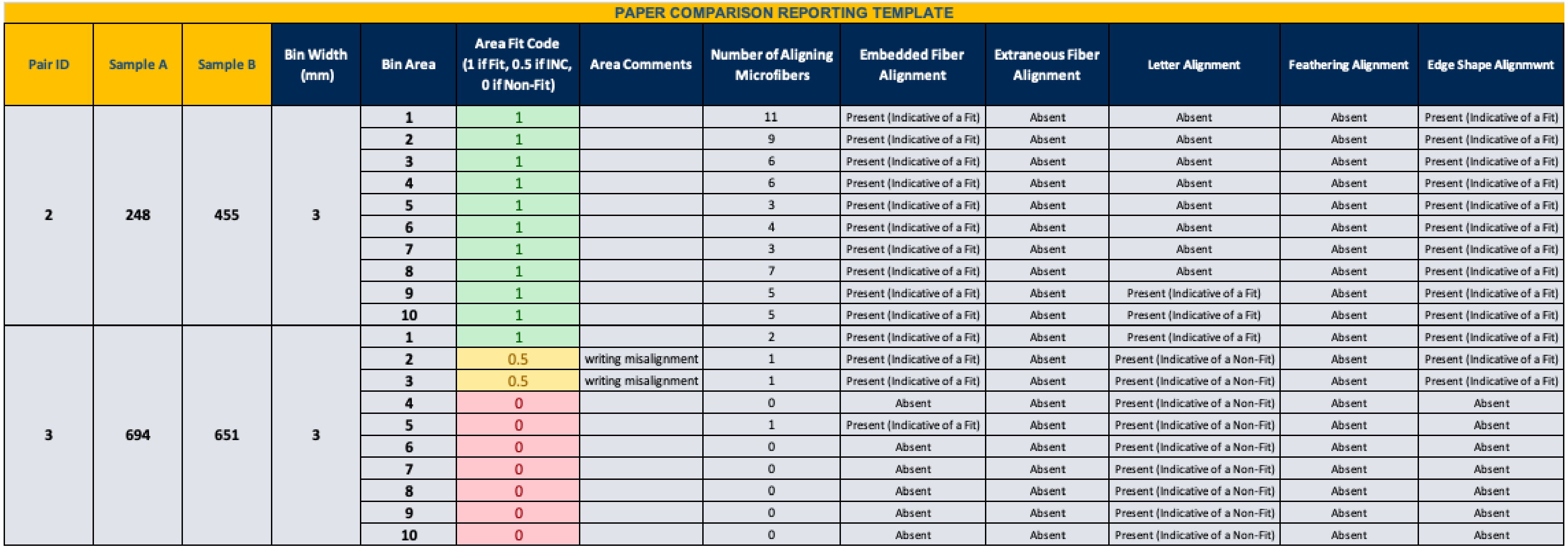
Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott Street and then goes to Turin and Rome and will join Colonel Parry and arrive at Athens, Greece, November 27 or December 2. Letters there should be addressed King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the 'Y. X.' Express tonight.

Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott Street and then goes to Turin and Rome and will join Colonel Parry and arrive at Athens, Greece, November 27 or December 2. Letters there should be addressed King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the 'Y. X.' Express tonight.

Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott Street and then goes to Turin and Rome and will join Colonel Parry and arrive at Athens, Greece, November 27 or December 2. Letters there should be addressed King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the 'Y. X.' Express tonight.

Printed Paper Sampling Template Example (not to scale). The grid blue lines are for demonstration purposes only and indicate the areas where the letter was cut or torn.

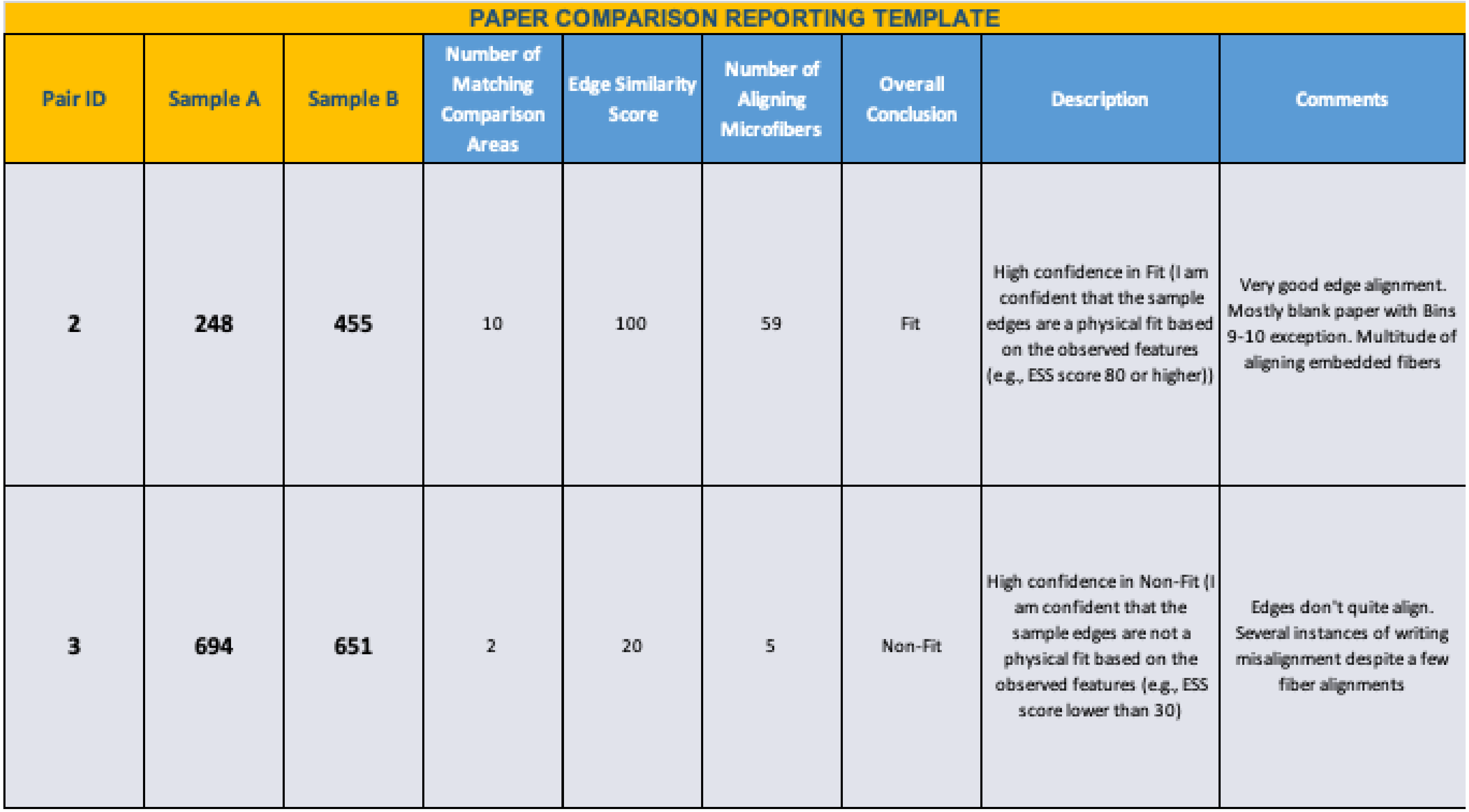
Paper Physical Fit Documentation Template Example (Part 1)



PAPER COMPARISON REPORTING TEMPLATE

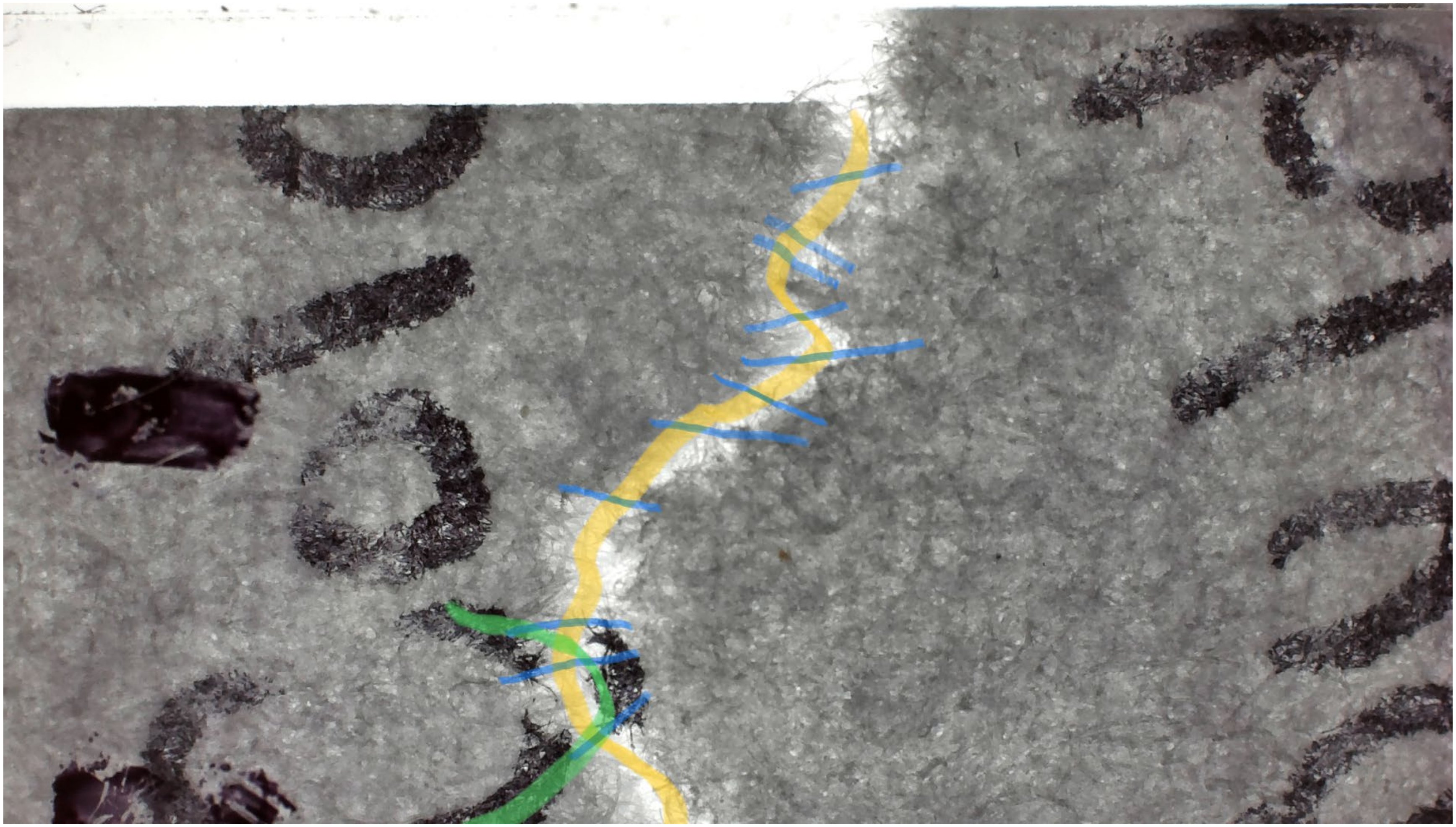
Pair ID	Sample A	Sample B	Bin Width (mm)	Bin Area	Area Fit Code (1 if Fit, 0.5 if INC, 0 if Non-Fit)	Area Comments	Number of Aligning Microfibers	Embedded Fiber Alignment	Extraneous Fiber Alignment	Letter Alignment	Feathering Alignment	Edge Shape Alignmwnt
2	248	455	3	1	1		11	Present (Indicative of a Fit)	Absent	Absent	Absent	Present (Indicative of a Fit)
				2	1		9	Present (Indicative of a Fit)	Absent	Absent	Absent	Present (Indicative of a Fit)
				3	1		6	Present (Indicative of a Fit)	Absent	Absent	Absent	Present (Indicative of a Fit)
				4	1		6	Present (Indicative of a Fit)	Absent	Absent	Absent	Present (Indicative of a Fit)
				5	1		3	Present (Indicative of a Fit)	Absent	Absent	Absent	Present (Indicative of a Fit)
				6	1		4	Present (Indicative of a Fit)	Absent	Absent	Absent	Present (Indicative of a Fit)
				7	1		3	Present (Indicative of a Fit)	Absent	Absent	Absent	Present (Indicative of a Fit)
				8	1		7	Present (Indicative of a Fit)	Absent	Absent	Absent	Present (Indicative of a Fit)
				9	1		5	Present (Indicative of a Fit)	Absent	Present (Indicative of a Fit)	Absent	Present (Indicative of a Fit)
				10	1		5	Present (Indicative of a Fit)	Absent	Present (Indicative of a Fit)	Absent	Present (Indicative of a Fit)
3	694	651	3	1	1		2	Present (Indicative of a Fit)	Absent	Present (Indicative of a Fit)	Absent	Present (Indicative of a Fit)
				2	0.5	writing misalignment	1	Present (Indicative of a Fit)	Absent	Present (Indicative of a Non-Fit)	Absent	Present (Indicative of a Fit)
				3	0.5	writing misalignment	1	Present (Indicative of a Fit)	Absent	Present (Indicative of a Non-Fit)	Absent	Present (Indicative of a Fit)
				4	0		0	Absent	Absent	Present (Indicative of a Non-Fit)	Absent	Absent
				5	0		1	Present (Indicative of a Fit)	Absent	Present (Indicative of a Non-Fit)	Absent	Absent
				6	0		0	Absent	Absent	Present (Indicative of a Non-Fit)	Absent	Absent
				7	0		0	Absent	Absent	Present (Indicative of a Non-Fit)	Absent	Absent
				8	0		0	Absent	Absent	Present (Indicative of a Non-Fit)	Absent	Absent
				9	0		0	Absent	Absent	Present (Indicative of a Non-Fit)	Absent	Absent
				10	0		0	Absent	Absent	Present (Indicative of a Non-Fit)	Absent	Absent

Paper Physical Fit Documentation Template Example (Part 2)

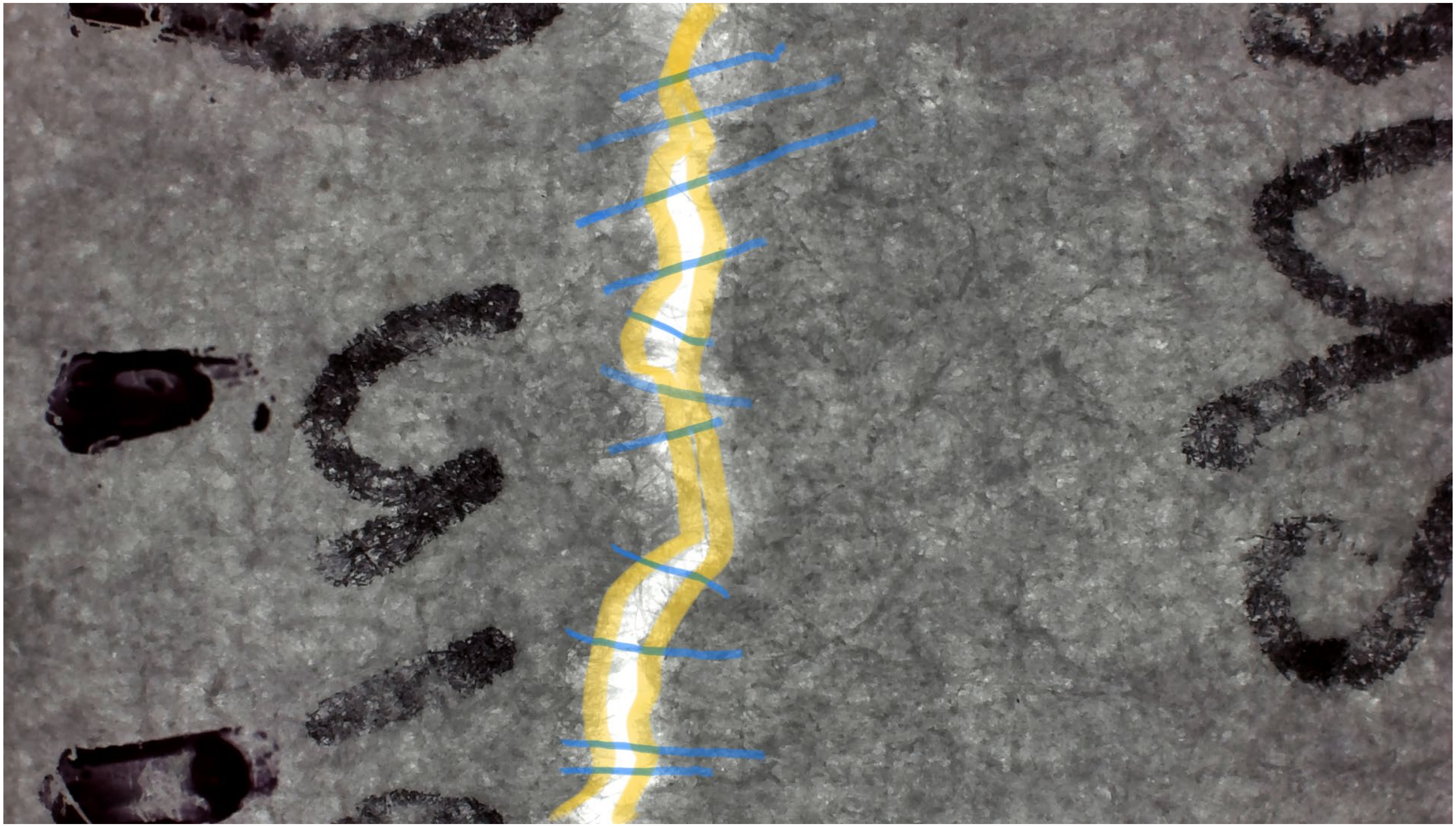


PAPER COMPARISON REPORTING TEMPLATE

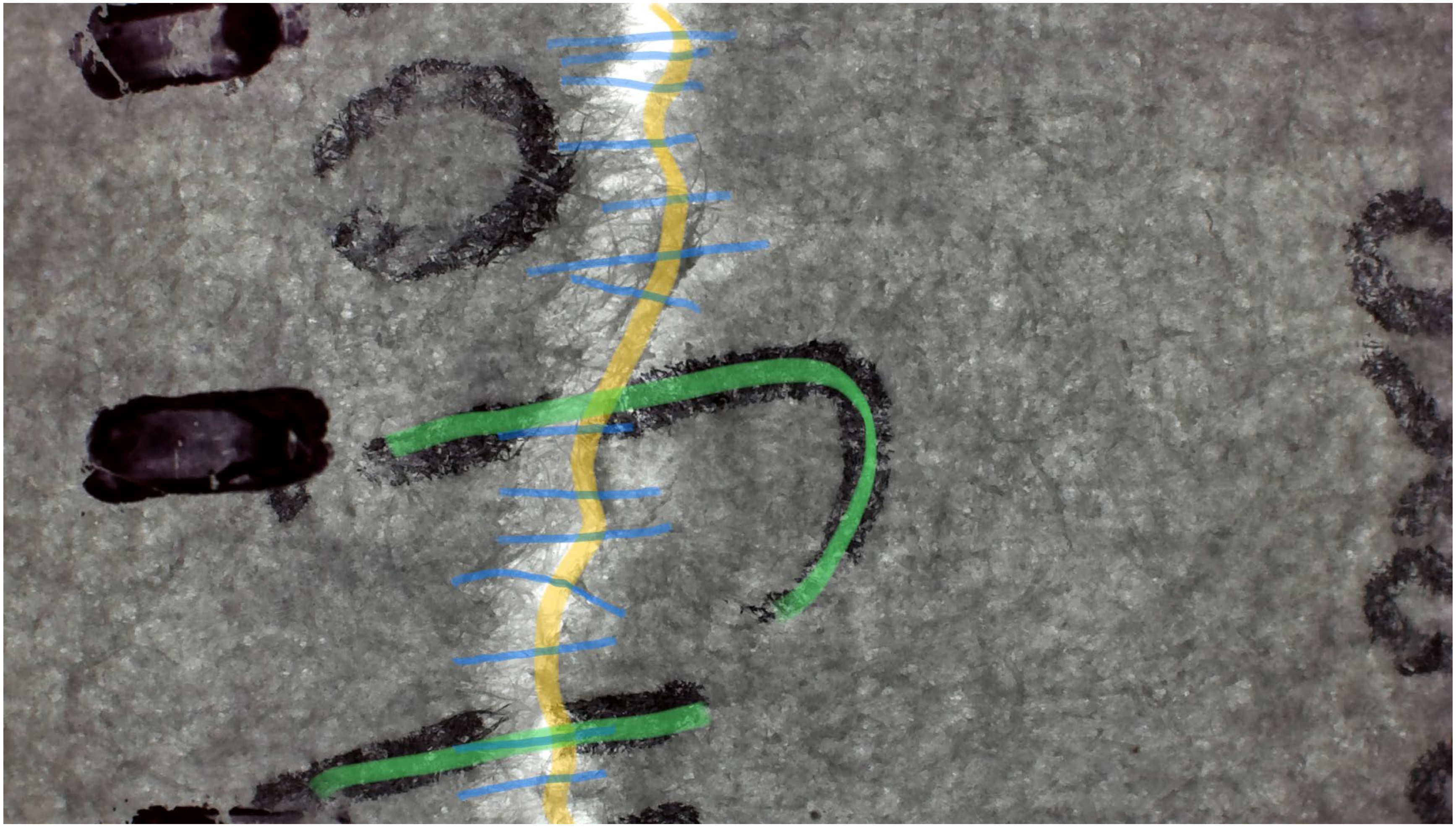
Pair ID	Sample A	Sample B	Number of Matching Comparison Areas	Edge Similarity Score	Number of Aligning Microfibers	Overall Conclusion	Description	Comments
2	**248**	**455**	10	100	59	Fit	High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed features (e.g., ESS score 80 or higher))	Very good edge alignment. Mostly blank paper with Bins 9-10 exception. Multitude of aligning embedded fibers
3	**694**	**651**	2	20	5	Non-Fit	High confidence in Non-Fit (I am confident that the sample edges are not a physical fit based on the observed features (e.g., ESS score lower than 30)	Edges don't quite align. Several instances of writing misalignment despite a few fiber alignments



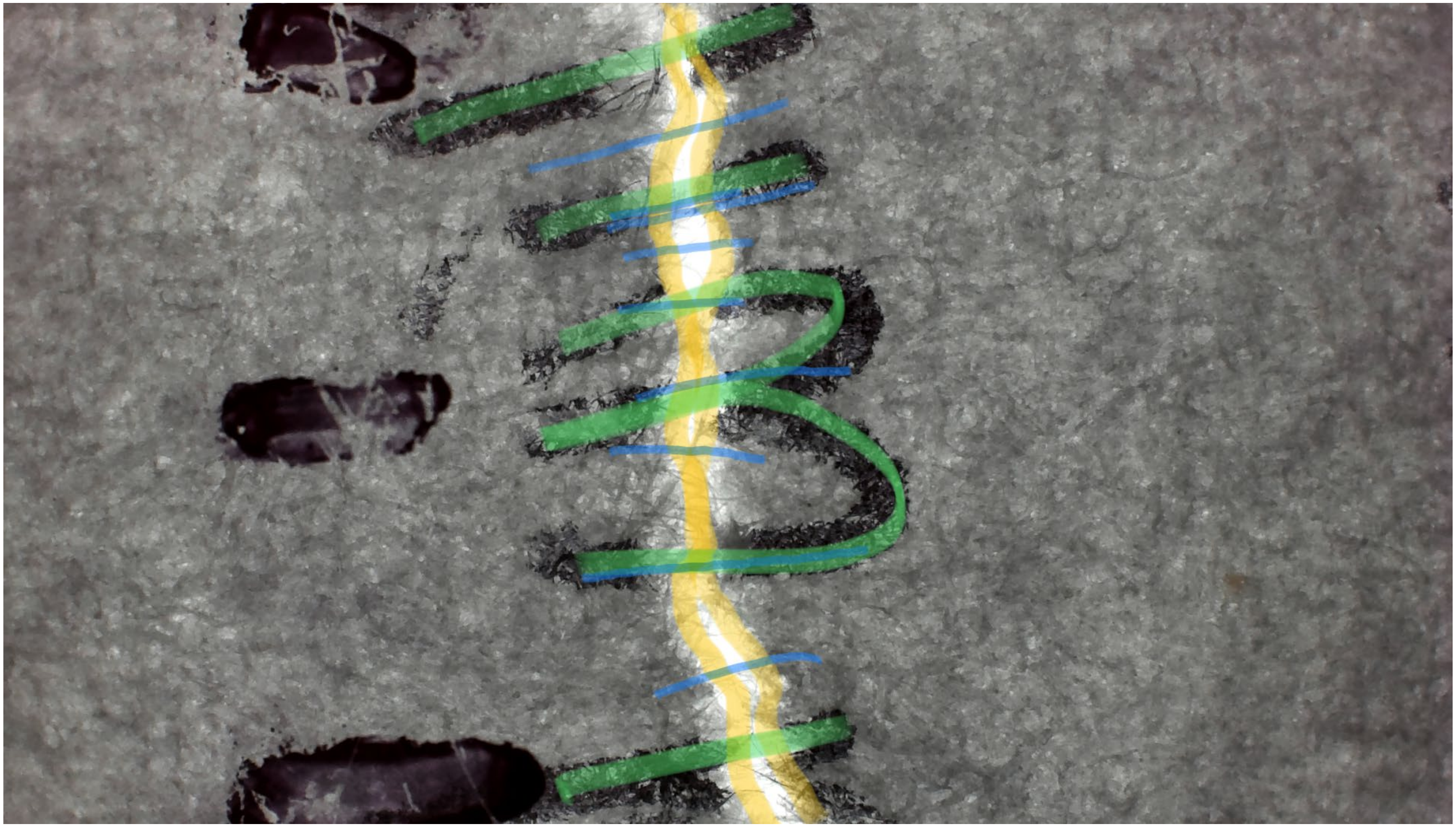
Microfiber Alignment Documentation of Bins 1-2 of Sample HWHT_1310_1321



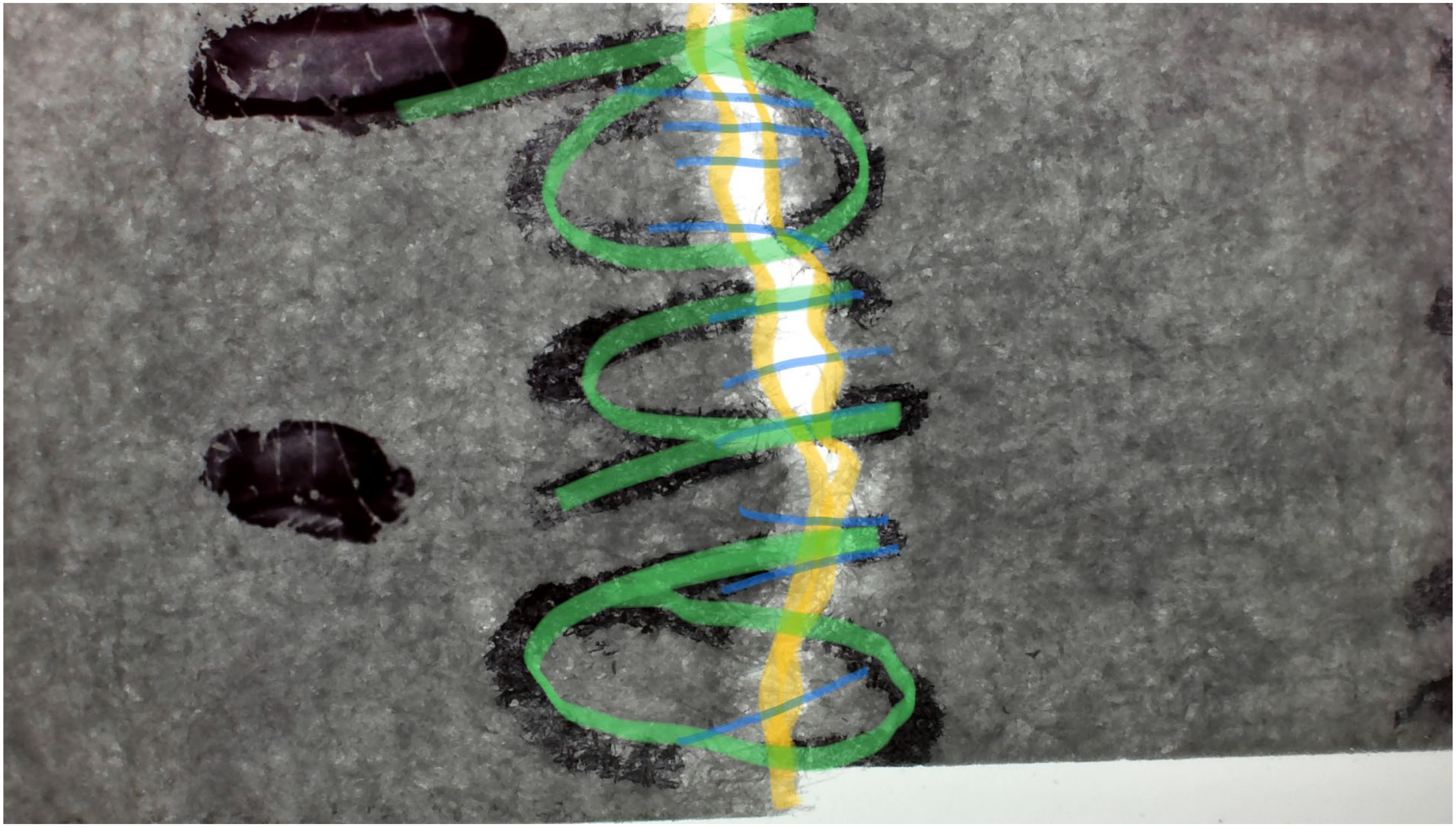
Microfiber Alignment Documentation of Bins 3-4 of Sample HWHT_1310_1321



Microfiber Alignment Documentation of Bins 5-6 of Sample HWHT_1310_1321



Microfiber Alignment Documentation of Bins 7-8 of Sample HWHT_1310_1321



Microfiber Alignment Documentation of Bins 9-10 of Sample HWHT_1310_1321

APPENDIX III. CHAPTER FIVE SUPPLEMENTAL MATERIAL

Postage Stamp Physical Fit Documentation Template Example

Pair #	Comparison Pair		Orientation of Fracture	Overall Comparison Edge Ends	Comparison Edge Qualifier	Unit Comparison Edge Ends	Weighting Factors (N=Number of Present Features)		Location of Microfiber Alignment (1=Present, 0=Absent)	Microfiber Location Verbal Descriptor (Crest/Trough in relation to lefthand sample)	Edge Comparison Comments	Edge Similarity Score (ESS) (macro)	Total Number of Microfibers in Alignment	Match Score 2^n n=total number of microfibers in alignment
	A	B	Vertical or Horizontal	(1=M, 0=NM, INC)	(M+, M-, IN, NM-, NM+)	Macro Feature Alignment (1=M, 0=NM)	Presence of Rare Features (Print Defects)	Microfiber Alignment						
1	PS-GR-11-E	PS-GR-12-W	Vertical	1	M+	1	0	1	1	upper crest	Overall macro alignment and agreement of micro features	100	34	17179869184
						1	0	2	1	upper crest, trough				
						1	0	3	1	upper crest, trough, lower crest				
						1	0	4	1	upper crest trough x2, lower crest				
						1	0	2	1	upper crest, trough				
						1	0	3	1	trough, lower crest x2				
						1	0	2	1	upper crest, lower crest				
						1	0	1	1	trough				
						1	0	2	1	upper crest x2,				
						1	0	1	1	upper crest				
						1	0	2	1	upper crest, trough				
						1	0	3	1	trough x3				
						1	0	8	1	upper crest, trough x2, lower crest x5				
2	PS-GR-17-E	PS-GR-35-W	Vertical	1	M-	1	0	1	1	upper crest	Overall macro edge alignment. Good agreement of micro features, but some possible continuations are not visible	100	13	8192
						1	0	1	1	upper crest				
						1	0	3	1	upper crest, lower crest x2				
						1	0	1	1	upper crest				
						1	0	2	1	upper crest x2,				
						1	0	1	1	lower crest				
						1	0	0	0					
						1	0	0	0					
						1	0	1	1	upper crest				
						1	0	0	0					
						1	0	1	1	upper crest				
						1	0	1	1	upper crest				
						1	0	1	1	lower crest				
3	PS-GR-110-S	PS-GR-319-N	Horizontal	1	M-	1	1	1	1	upper crest	Grime in Bin 1, Bin 12 very large/thick fiber, macro level alignment	100	10	1024
						1	0	1	1	trough				
						1	0	0	0					
						1	0	0	0					
						1	0	0	0					
						1	0	2	1	lower crest x2				
						1	0	0	0					
						1	0	0	0					
						1	0	1	1	trough				
						1	0	3	1	`				
						1	0	1	1	trough				
						1	0	1	1	trough				
						1	0	0	0					