

CONSISTENCY OF JUDGMENTS OF A PROFESSIONAL PANEL AND WEARERS
REGARDING CHARACTERISTICS OF SELECTED SLEEPWEAR FABRICS

by

Margaret Jane Gallagher

Thesis submitted to the Graduate Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Clothing, Textiles and Related Art

APPROVED:

M. A. Zentner, Chairman

W. L. Mauldin

J. E. Mann

R. S. Merkel

December, 1978

Blacksburg, Virginia

ACKNOWLEDGMENTS

The author would like to express sincere appreciation to Dr. Mary Ann Zentner, Head of the Department of Clothing, Textiles and Related Art, for her role as committee chairman. Her encouragement, understanding and guidance were invaluable to the completion of this project. Thanks are also extended to committee members Dr. William Mauldin, Professor, Clothing, Textiles and Related Art and Dr. Robert Merkel, Assistant Professor, Clothing, Textiles and Related Art for their assistance and support. A special debt of gratitude is owed to Dr. Jerry Mann, Assistant Professor, Statistics, for his patience and countless hours of work adapting a relatively new statistical procedure to this application.

Appreciation is also extended to the individuals who participated in the wear study and to the professional panel members:

, , and Dr. Mary Ann Zentner.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	ii
LIST OF TABLES	v
Chapter	
I. INTRODUCTION	1
II. REVIEW OF LITERATURE	5
Laboratory Panel Observations	5
Panel and Consumer Evaluations	13
Sleepwear Fabric Evaluations	21
Summary	25
III. STATEMENT OF THE PROBLEM	27
Objectives	28
Hypotheses	28
Assumptions	29
Limitations	29
Operational Definitions	30
IV. PROCEDURE	31
Selection of Test Fabrics	31
Sample Selection	31
Data Collection	32
Wearer Evaluation	33
Panel Evaluation	33
Statistical Analysis	35

Chapter	Page
V. FINDINGS AND DISCUSSION	39
Comparison of Judges	42
Comparison Between Judges and Wearers	62
VI. SUMMARY	84
Category I Hypothesis	85
Category II Hypothesis	88
Recommendations for Future Research	93
REFERENCES	93
APPENDICES	97
A. WEARER RATING CHECK SHEET	98
B. PROFESSIONAL RATER CHECK SHEET	100
VITA	103
ABSTRACT	

LIST OF TABLES

Table		Page
1.	Weights Assigned to Calculate Proportion of Observed Agreement	38
2.	Contingency Table for Judges I & IV for Hand of Series 100	40
3.	Contingency Table for Judges I & III for Shape Retention for All Series	51
4.	Weighted Kappa and Z Scores of Agreement Between Judges on Performance Factors of All Series Fabrics	63
5.	Weighted Kappa and Z Scores of Agreement Between Judges and Wearers on Per- formance Factors of All Series Fabrics	69

:

Chapter I

INTRODUCTION

It is important that performance characteristics of textile products be evaluated to maintain quality and provide information to industry and consumers. Consumer satisfaction is met by designing fabrics that meet the requirements of a number of specific end uses. To determine fabric characteristics, physical laboratory testing and/or subjective visual inspections of fabric wear may be made after laundering or actual consumer wear. A substantial amount of work has been done in determining the degree to which objective tests predict actual wear. Little has been done, however, to determine how well panels of human evaluators can predict consumer satisfaction with textile product performance. While these subjective evaluations are necessary in the evaluation of many textile properties, the procedure is based largely on unproven assumptions.

In a paper presented to the New York Chapter of the American Association of Textile Technology, Whitcomb suggested that perhaps the consumer did not notice wear that researchers found objectionable. She noted that very little had been done to determine whether or not this was true (1969, p. 67). Powderly stated that home economists were more critical of the performance of men's shirts than were consumers (1970, p. 90). He observed, however, that

with the rising educational levels of today's women and high school courses in consumerism, homemakers were becoming more aware of textiles. "The requirements of tomorrow's consumer will be those of today's home economists. Their standards will be higher and their demands for esthetics more critical" (Powderly, 1970, p. 90). To date, there is no conclusive body of knowledge to confirm or deny these statements.

Discussing the development of wash-and-wear performance standards in 1960, Bercaw stated, "We realize that the eyes and mind of the consumer are the ultimate judge in assessing appearance of fabrics and garments" (p. 39). He added that "as . . . differences in appearance become more and more subtle (they are) difficult to detect consistently on the basis of subjective evaluations" (pp. 39-40). Many improvements have been made in fabric durability and appearance characteristics in recent years. Thus, textile experts believe that consistency of subjective evaluations is increasingly important. As the differences become less distinct, professionals may or may not evaluate fabrics in the same manner as consumers. Thus, whether or not trained persons can accurately predict consumer attitudes should be of concern to those involved in the marketing of textile products. The ultimate aim of industry is to give consumers the level of performance they desire.

The cost of attaining satisfactory performance levels has been of some concern in recent years. Discussing consumer standards,

Powderly emphasized this point, "It costs money to get performance and no one will thank you for selling them something they don't want. This makes it vital to know what the consumer wants and doesn't want" (1970, p. 89). Wear trials have been seen as a meaningful predictor of fabric performance (Wright & Clary, 1969). The drawbacks to this method are obvious, however. The cost of a wear study is great, as is the time involved to obtain results. After the in-service period, garments may be evaluated by wearers or a trained panel or both. The number of persons composing trained panels can vary. The AATCC Technical Manual recommends three persons for the majority of test methods involving subjective evaluations. If the number of persons evaluating worn garments were reduced, part of the expense of these studies could be eliminated. The degree of agreement between panel members could confirm or deny the need for a number of panel members. If all members evaluate in the same manner, the use of only one evaluator might be sufficient. The trained panel evaluation could be eliminated in some cases if it were evident that professional raters made the same observations that consumers did.

There appears to be little information in the literature concerning the consistency of panel ratings or of ratings between panels and consumers. In numerous studies, ratings of persons were merely averaged, with no attempt to correlate ratings. It is the purpose

of this study to determine whether trained panel members are consistent in their evaluations of fabric performance and whether a panel will accurately predict consumer reactions to selected fabrics. Data were obtained as part of the Northeastern Regional Research Project - 79, Consumer Market and Laboratory Studies of Flame Resistant Textile Items. Funding was obtained from the United States Department of Agriculture through the Research Division of Virginia Polytechnic Institute and State University.

Chapter II

REVIEW OF LITERATURE

Subjective evaluations of fabric performance have been carried out in many research studies. The review of literature will include references to visual evaluations of fabrics suitable for various end uses. Three areas will be covered: evaluations by laboratory panels, evaluations by both a panel and consumers, and evaluations of sleepwear fabrics. The latter are included because the data used in the current work were obtained during the course of a wear study on women's sleepwear.

Laboratory Panel Observations

Subjective evaluations were used by Erickson (1971) to determine appearance characteristics of 65/35% polyester-cotton shirts with permanent press and soil release finishes. Six men wore the shirts for either 12 or 36 laundry periods. Three trained observers then evaluated the shirts using AATCC Test Method 124-1969 (1). They rated fabric smoothness on a six-point scale from excellent to unacceptable. The researcher perceived the need for more objective controls for the subjective evaluations. She stated that "the trained observers were influenced by their own personal backgrounds in regard to the ratings they gave the shirts" (p. 46). The development of additional standards was suggested as an aid to this problem (p. 48).

Several other researchers report variations in panel member ratings. In one study, Christian (1972) investigated the effects of laundering and drycleaning on performance characteristics of a foam laminated fabric. A 100% acrylic face fabric with a 100% acetate tricot backing was rated subjectively for visual appearance and four characteristics of hand: pliability, softness, stretch and spring. Three graduate students rated hand in accordance with ASTM Method D-123-64 and appearance based on AATCC Test Method 124-1969 (pp. 29-30). Overhead fluorescent lighting was used in the judging area.

The mean sources of the ratings were reported. The pliability ratings fluctuated from the fifth to the fifteenth laundry intervals. The fabric was rated more pliable after the tenth laundering than after either the fifth or fifteenth intervals. The researcher explained that "possibly the method of evaluation was not sensitive enough to take care of the subjective reasoning of the evaluators. The raw data . . . shows scores were sometimes very extreme" (p. 87). The evaluators' ratings varied from a difference of zero to two points, on a five-point scale, for softness and pliability of laundered and drycleaned specimens (pp. 135-138).

Inconsistencies of evaluations of performance were found in wear testing of fabrics done by the United States Army (Murdock, 1969). Following an accelerated wear period, the garments were evaluated by two subjective methods--the GETA System and the British Scoring

System. Concern about the subjective nature of these evaluation procedures was expressed (p. 20). Both subjective systems required a count and charting of evidences of wear. The failures recorded using the GETA method were: wear areas, holes, frays, tears, broken stitching and bar tacks, and button failure. The British Scoring System identified only tears and holes as failures.

A test was conducted in which over 400 holes were charted by these two methods. A 37% inconsistency in recording the failures was reported using the procedures (p. 21). An objective measuring device, using Beta radiation to detect fabric weight changes during wear, was proposed by Murdock as a more accurate means of evaluation. The results obtained using the instrument were found to be more reproducible than the subjective rater evaluations (p. 21).

Fabric appearance, ease of care, durability and effects of sewing methods on seam appearance of four polyester-cotton blends were evaluated in a study by Tyler (1972). Included in the study were 50/50% Fortrel polyester-cotton, 50/50% Kodel polyester-cotton, 65/35% Dacron polyester and 80/20% Dacron polyester-cotton, all having a permanent press finish. Seam appearance was evaluated by a three-member panel of experienced judges at specified laundry intervals. Appearance was judged according to modified AATCC Test Method 124-1969, procedure IIB. The AATCC Three Dimensional Durable Press Replicas

were used as standards. Both seam appearance and fabric appearance were inspected during daylight hours, with shades half drawn. Overhead fluorescent lighting was used.

Mean ratings of the judges provided appearance scores for the fabrics. Although no tests of reliability appear to have been done, the researcher observed that the analysis of data was affected by rater variation. One judge generally rated the fabrics one class lower than did the other two evaluators, thus lowering the average scores. The researcher noted that, because of the variability of scores, the mean was not a good indicator of seam appearance (p. 23).

Headrick (1968) noted a variation in researcher and panel ratings of the performance of white table napkins. Those with a durable press finish and others with a soil release durable press finish were evaluated at certain service and laundry intervals by the experimenter and a panel of untrained judges. The judges rated the napkins, in a booth illuminated by three fluorescent lights, on the properties of surface smoothness, change in color, hand and fibrillation. Surface smoothness was evaluated with AATCC Photographic Standards for Wash and Wear for comparison. The AATCC Gray Scale for Evaluation of Color Change and an unused napkin were used as standards for rating color change. Hand and fibrillation were rated on five-point scales.

Means of the judges' ratings were used for analysis. The researcher detected a discrepancy in her assessment of color change and

the evaluation by the panel. She observed that the judges rated color as excellent, but stated, "Although untrained panel members did not detect large differences in color change in judging under fluorescent lighting, the experimenter felt that greater differences actually existed between the two fabrics" (p. 68). It was reported that these differences in color were more evident to the experimenter than to the panel.

When numerous performance factors were assessed, ratings sometimes varied and sometimes were in agreement, depending on the factors evaluated. Freeman (1972) found differences of agreement in evaluating the effectiveness of stain release finishes on white, 50/50% polyester-cotton fabrics. Stains consisted of "Nujol" oil, blue tempera paint and coffee with cream and sugar. The researcher and three trained panel members rated laundered stained samples after 1, 10, 20 and 30 launderings. The visual characteristics of laundered samples were inspected using the "at-a-glance" method in comparison with Stain Release Replicas for oily stains developed by Deering Milliken Research Corporation and Stain Release Replicas for pigmented stains developed at Southern Illinois University. The procedure for evaluating the fabrics was adapted from AATCC Test Method 139-1970. Direct overhead fluorescent lighting was used (p. 23). Freeman found that her ratings agreed closely with overall panel ratings. It was noted,

however, that panel ratings were slightly lower than the researcher's for tempera paint and coffee stains. No difference was observed between the ratings of the oil stains.

Two fabrics of 50/50% polyester-cotton, manufactured by different companies and treated with different types of optical brighteners, were evaluated after soiling and laundering by Perenich (1972). After varying amounts of exposure and dark storage, the appearance of the samples was rated by a panel of four persons from a department of clothing and textiles. Overhead fluorescent lighting was used in the evaluating situation. Soiled and unsoiled samples were used as standards for comparison. Each panel member rated the samples twice. The researcher indicated that this procedure was carried out to check reliability (p. 30). The two scores were averaged to obtain a visual rating for each judge. The experimenter observed a high degree of agreement among the ratings of panelists for all samples except those subjected to dark storage. These ratings were significantly different among the judges (p. 64).

Close agreement among panel members has been found in several studies. Booker (1970) reported rater agreement in assessing subjective panel evaluations of select performance characteristics of stretch and conventional woven fabrics. Two pairs of slacks of 62/35/3% Trevira polyester-Avril rayon-Lycra spandex and 65/35% Dacron polyester-Avril rayon were worn by six participants for up to 30 wash periods.

Three judges rated durable press characteristics according to AATCC and ASTM testing procedures (p. 67). The researcher indicated that judges were consistent in their assessment of surface appearance (p. 68).

Ruppenicker, Rhodes, Markezich and Little (1972) found high correlations among judges' ratings, indicating close agreement among the raters. Edge abrasion of durable press fabrics which were made into cuffed trouser legs was evaluated by five graders. Four cotton fabrics were treated with three different durable press finishes (p. 33). Ratings of crease damage to trousers which were laundered 40 times were made according to the AATCC procedure. Inspections took place in a darkened room, using a frosted light source. Then photographic standards for edge abrasion were used as comparisons when rating the fabrics. Agreement between the five graders was analyzed statistically by correlating judges' ratings for each fabric. Correlation coefficients varied from .95 to .98, indicating close agreement among the raters (p. 35).

Comparisons of judges' ratings are seldom made in textile studies. Larson (1973), Bostrom (1973) and Suansamroeng (1974) are among those who have used panel evaluations without determining the consistency of their ratings.

Nine types of men's work trouser fabrics were evaluated through 30 home launderings by Larson (1973). Fabrics included were 100% cotton and 87/13, 75/25, 63/37 and 50/50% of both cotton-polyester

and the same blends of cotton-high modulus polyester. Two performance characteristics that were subjectively rated were durable press appearance and colorfastness. A three-member panel evaluated fabric smoothness according to AATCC test method 124-1969 (2b) (p. 27). Two eight-foot Type F96 Cool White fluorescent lights hung above and in front of fabrics being inspected. Colorfastness was also rated by a three-member panel using the AATCC Gray Scale for Evaluating Change in Color (2a) as a standard (p. 29). The ratings of the judges were averaged for statistical analysis. Consistency of the ratings was not reported.

Durable press appearance and pilling propensity of all-cotton and polyester-cotton blend sheets, after 50 use periods, were subjectively evaluated in a study by Bostrum (1973). A three-member panel rated durable press appearance on a five-point scale using the AATCC Three Dimensional Durable Press Replicas as guides. The ratings of the judges were averaged to obtain a rating for each sheet. One person rated pilling by a count of pills in a specific area. Room conditions and type of lighting were not stated. No tests of rater consistency were reported in this research.

Knit shirts of 100% cotton, 100% Arnel triacetate and 65/35% Arnel triacetate-Dacron polyester were worn and laundered, and laundered but not worn, in the study by Suansamroeng (1974). Appearance was measured at specific laundry intervals according to AATCC Test Method

124-1969 (1c), modified to use two observers instead of the three recommended in the test procedure (p. 30). Seam appearance was rated in accordance with AATCC Test Method 88B-197 (1b), using only two trained observers (p. 31). One person rated color retention according to AATCC Test Method 36-1972 (1a). Color change was compared with a control garment using the AATCC Gray Scale for Color Change as a guide (p. 33). Pilling was measured by a count of pills in the collar area. When two persons inspected garments, the scores were averaged for analysis. No comparison of judges was made.

Panel and Consumer Evaluations

A pattern of agreement or disagreement between panel and consumer ratings could not be detected in the literature. When several characteristics were considered by the two groups, agreement sometimes varied with the characteristic evaluated.

This variation was evident in a study by Morris, Margerum and Young (1972). One hundred and twenty men's durable press shirts of either 100% cotton or 65/35% polyester-cotton were worn by 60 men as part of a United States Department of Agriculture Western Regional Research Laboratory Project. Twelve shirts were laundered but not worn. The garments were evaluated after 12 and 24 wear and laundry periods at each of five laboratories. Subjective evaluations of durable press performance were made according to AATCC Test Method 124-1967, using AATCC plastic replicas for fabric smoothness as

standards. Pilling was evaluated by three judges at each laboratory, but standards of comparison were not used. The raters reported the existence or absence of pills (p. 40). Wearers rated the shirts on whiteness, fabric smoothness, need for pressing, abrasion, pilling and shirt performance.

Mean values of the judges' ratings were recorded. The researchers found that participants and raters agreed in their assessments of fabric smoothness. Both groups felt the polyester-cotton fabric had better appearance than did the cotton fabric. There was, however, a discrepancy in the observations of pilling. Only 18% of the wearers reported pilling on the polyester-cotton shirts, in contrast to 55% of the judges who noted that factor and were more discriminating. Neither group reported pilling on the 100% cotton shirts (p. 42).

A wear study of blouses of a 65/35% polyester-cotton permanent press fabric used subjective ratings by wearers and a panel of judges to evaluate fabric performance. Characteristics that were subjectively rated included surface smoothness, seam puckering, crease retention, color change, hand and fibrillation (Harper, 1966). As in the study by Morris et al. there was agreement between the groups on some factors, but the panel was more critical on one factor.

The wearers rated the blouses after 0, 5, 15 and 25 launderings, using a three-point scale of good, fair or poor, on the above appearance factors. Harper noted that the ratings were highly subjective.

The panel, consisting of ten faculty and graduate students in a textiles and clothing department, evaluated surface appearance, seam puckering and crease retention in relation to AATCC Photographic Standards for permanent press. Color change was rated using the AATCC Gray Scale for Color Change as the standard. Hand and fibrillation were rated on five-point scales as appropriate. Evaluation of all factors, except color change, was made using fluorescent ceiling, desk and flood lights. Color change was rated in natural light of northwest exposure (p. 58). Harper stated that these ratings were also subjective, but probably less so than wearer ratings (p. 58).

Findings revealed that raters were within one interval of each other, on a five-point scale, when judging all performance characteristics except crease retention (p. 87). Evaluations spread over three intervals for that factor. Panel members consistently gave high ratings for surface smoothness (p. 86). Because wearers and raters made evaluations on different scales, it is very difficult to visually compare panel and wearer ratings. Harper observed, however, that seam puckering was more evident to the panel than to wearers (p. 89). Consistency of panel members and wearers in rating surface smoothness was noted (p. 89).

No pattern of consistency among raters was found by Ball (1971). The appearance of men's khaki durable press trousers of five fabric-fiber-finish combinations was measured during 30 wash-wear cycles,

using wearer and trained panel observations. White-collar and blue-collar workers wore the slacks. Some were retained in the laboratory and laundered. A panel, composed of three textile technologists, evaluated fabric smoothness, seam smoothness and crease sharpness at specific intervals, using the appropriate AATCC standard procedures and photographic standards. They also evaluated these characteristics while observing the trousers on the wearers and using photographic standards developed by the researcher. Wearers were asked to give their opinions of fabric smoothness, seam smoothness, crease sharpness and comfort at regular intervals.

Mean ratings of the panel and both wearer groups were compared. The white-collar group gave a higher rating for fabric smoothness than did the panel. The same was true for the blue-collar group. The panel rated the trousers higher using the AATCC standards than with the garments on the men (p. 71). Both groups rated seam smoothness lower than the panel ratings. Use of the AATCC standards by the panel produced lower ratings than the evaluations of garments on the men (p. 94). The blue-collar and white-collar groups rated crease retention higher than panel ratings. Trouser evaluations on the men and using AATCC standards were the same for crease retention (p. 86).

A study of disposable laboratory coats involved in-service testing and subjective and objective evaluations to assess performance (Yuan, 1967). Twenty laboratory coats were worn by four subjects

for approximately 40 hours per coat and rated by the researcher and a three-member panel trained to judge selected performance characteristics. They evaluated appearance and signs of wear on a three-point scale using standard textile test methods (p. 23). Frequency counts of responses were reported. Three of the subjects found the appearance of the coats acceptable and one rated appearance as unsatisfactory (p. 59). All panel members noted broken stitches and enlarged needle holes in worn coats. The panel rating of color change was consistent with the instrumental measurements. Panel, researcher and participants were in agreement in rating the hand of garments as less stiff after wear (p. 62).

Agreement between consumer and panel ratings was observed in several studies. In one study, Bradley (1968) conducted a wear trial and visual evaluations to determine the performance of shirts of 65/35% polyester-cotton with either durable press finish, durable press and soil release finish or no durable press treatment. After stated laundry periods, a panel of three trained persons made at-a-glance inspections of the shirts under a single forty-watt ceiling light. Seam puckering was evaluated using AATCC standard photographs. Retention of original appearance was rated on a three-point scale and mean scores were reported. Wearers were instructed to rank shirts by preference at each laundry period. After the final laundering, they were asked to indicate acceptability of the appearance of the shirts.

The panel and wearers were in agreement in ranking the untreated fabric highest for appearance. Regarding the consistency of the panel and wearers overall, the experimenter states, "The panel and wearers were in close agreement at all times" (p. 68).

Chamberlain (1972) investigated the performance of 100% Qiana nylon knit sport shirts. After an in-service wear period, wearers and a three-member panel inspected the garments. Appearance on the wearers was evaluated by both the panel and participants, using an original for comparison (p. 2). The panel rated colorfastness by comparing the shirts to the control and the AATCC Gray Scale for Color Change.

The three panel members were consistent in rating color slightly lighter than the original. Both participants and panel rated garment construction as satisfactory. The researcher's findings were also in agreement with the panel for that factor (p. 37).

Subjective evaluations by wearers and five judges followed wear and laundering of durable press dress shirts of 100% cotton, 35/65, 50/50 and 65/35% polyester-cotton (Ogle, 1968). Wearers rated two randomly selected shirts at specified laundry intervals on surface smoothness, seam puckering and appearance of collars. The judges rated the same shirts on surface smoothness, seam puckering, color change and fibrillation. Although the consistency of panel judgements was not determined, it was indicated that wearers and judges agreed

in their assessment of whiteness retention of the 100% cotton fabric (p. 66). Agreement between the wearers and panel was not indicated for the other fabric types.

Donnel (1965) investigated construction and care of deep pile fabrics, using an in-service wear test. Six overshirts of 100% acrylic deep pile fabric were constructed and four put into wear for 12 weeks. Two shirts were retained as controls. The garments were worn an average of 20 hours total and laundered weekly (p. 21). The evaluations were made by a trained panel and wearers, though the number of persons on the panel was not specified. The panel rated appearance, hand, seams and size and shape of worn garments relative to a new garment. No method of calculating scores was reported. Wearers visually inspected their shirts in relation to an original and rated appearance on a four-point scale. Consistency tests were not performed among the raters. Panel and wearer agreement regarding fabric softness after laundering was, however, observed by the researcher (p. 53).

Kitchen towels of three different constructions and varying percentages of cotton and rayon were tested in-service and rated by users and a three-member panel (Inlow, 1966). At-a-glance evaluations of appearance and colorfastness of the original and laundered towels were made by the trained panel. The AATCC Gray Scale for Evaluating Staining (ISO Recommendation R105/I, Part 3) was used for two types

of towels and the AATCC Geometric Gray Scale for Evaluating Color-fastness was used for the third type. Participants were asked to rank the towels by preference before laundering and after 50 wash cycles. One towel of each type was selected at random for visual evaluation by the panel. These were inspected in comparison to an original of each type under a 40-watt fluorescent ceiling light. The judges classified the towels for color retention (on a two-point scale) and acceptability for use, and ranked them in order of preference. The scores of the judges were averaged.

Inlow found that, although fading occurred, towels were acceptable to both consumers and the panel. It was noted that consumers did not detect the considerable shrinkage that had occurred. The panel and participants were found to be consistent in their ratings overall, however (p. 51).

Fargus (1972) conducted wearer and panel evaluations of 82/18% nylon-spandex stretch lace dresses. Five dresses were constructed and one was retained as a control. The remaining four were worn through 20 laundry periods. Wearers rated dress appearance after each laundering. At the completion of the wear trials, they rated the dresses by responding yes or no to visual change, physical change, comfort, construction failures, color change and shape retention. The panel, composed of three trained persons, made at-a-glance evaluations of the worn dresses in comparison with the original. Construction

and fit were observed and rated on a three-point scale. Standard overhead lighting was used for the inspection of garments. Because wearers and panel did not evaluate the dresses on the same performance characteristics, no comparisons of the two groups were made. Tests of consistency of the trained panel were not included in the study.

Sleepwear Fabric Evaluations

Few evaluations of women's sleepwear fabrics and wear studies of women's sleepwear have been reported in the literature. The larger investigation, of which this work was a part, was initiated to explore changes in both objective and subjective properties of several commercial sleepwear fabrics through laboratory testing and a wear study. Women's sleepwear was chosen because nightgowns were one of six leading categories of clothing proposed for federal flammability regulations.

One study of consumer acceptance of flame retardant children's sleepwear of summer weight fabrics was conducted at Winthrop College (Laughlin, 1976). Nightgowns in sizes 4, 5 and 6X were constructed of 65/35% polyester-cotton untreated fabric, 100% woven polyester with a flame retardant finish and a knit of 50/50% flame retardant cellulose triacetate-polyester (p. 51). Three hundred and twelve gowns made from these fabrics were worn by girls during a summer season.

At the end of the season, the girls' mothers indicated their preferences. Objective and subjective testing was also performed in the laboratory. A panel of three experienced observers evaluated all returned garments for appearance and pilling by inspecting the worn gowns alongside a new gown. Uniform lighting was used in the testing area (p. 53).

Panel ratings were averaged. Although statistical tests were not reported, it was observed that the panel and consumers differed in their opinions of the garments regarding appearance and pilling. The panel preferred the polyester over the untreated polyester-cotton fabric whereas the consumers ranked garments of those fabrics the same (pp. 37-38).

As a continuation of the above study, winter nightgowns made from seven fabrics, in sizes 7-14, were evaluated after wear by consumers and a trained panel (Laughlin, 1977). The fabrics included in the study were 50/50% polyester FR cellulose triacetate knit, 100% FR polyester knit, 80/20% acetate-polyester knit, two fabrics of 100% treated polyester flannelette, 65/35% cotton-polyester treated flannelette and 100% untreated brushed tricot nylon. The number of persons serving on the panel and room conditions for evaluations were not mentioned. The averages of the judges' ratings and consumers' ratings of appearance and pilling were given. Some

consistency between the panel and consumer rankings was reported for these factors, but statistical significance levels were not presented (pp. 73-74).

Four types of boys' ready-to-wear pajamas, each of a different style, fiber content and fabric construction, were used in a regional research study in a number of states by Avery, Schaeffer, Densmore, Ahrens, Mauldin, Peters, Quick, Smith and Zentner (1978). One hundred and twenty children wore a total of 264 pairs of pajamas during a seven-month period. Thirty-two pairs of unworn pajamas were laundered 50 times (p. 8). Visual evaluations were made by mothers of the wearers following the fifth, twenty-fifth and fiftieth home launderings. Satisfaction record cards were provided to the mothers to record their judgements of soil removal, colorfastness, shape retention, appearance, hand and durability. Definitions of these terms appeared on the back of each card for use by the participants. Researchers inspected the garments after the designated laundry intervals. Visual evidences of damage and wear, resulting in changes in appearance and hand, were reported.

After the fifth wear period, the judges rated one fabric type lower than did the participants on the factors of soil removal and colorfastness (p. 22-23). Both judges and homemakers rated another fabric type lowest for shape retention. Appearance of the garments was generally rated lower by judges than by the mothers after 25

laundryings (p. 23). After 25 wear periods, one fabric was given a high rating on hand by the professionals and the homemakers (p. 24). Another fabric was consistently rated highest for overall durability by both groups throughout all wear periods. Results for both groups were reported; however, no statistical tests of significance were given.

In another investigation, nightgown fabrics made from four different fiber combinations were evaluated by consumers after wear and laundering (Davis, 1978). Included in the study were 100% polyester brushed knit, 100% nylon tricot knit, 80/20% acetate-polyester brushed knit and 80/20% acetate-nylon brushed knit. Each of fifty-four women wore two gowns constructed from the fabrics for a period of eight months, laundering them at least 10 times. The wearers' satisfaction, preferences and care procedures were determined. Interviews were conducted before the wear period to obtain demographic data and usual laundry procedures. At the end of the wear period, subjects were interviewed regarding gown preference and asked to rate the gowns on selected performance characteristics.

Soil removal, colorfastness, appearance, hand and durability were all rated either excellent or good by over half of the wearers of all fabrics (pp. 47-48). The researcher reported that static electricity was a source of dissatisfaction and was a major problem with all of the gowns (p. 53).

Paek (1978) developed an index of variables related to fabric hand. Response profiles of three panels were obtained relative to the hand of selected flame-retardant sleepwear fabrics. The three panel groups consisted of 62 homemakers, 112 college students and 3 trained laboratory personnel. The response profile was determined using mean hand ratings of each group (p. 487). When the mean ratings of these three groups were compared for three tricot knit fabrics, the trained panel gave lower scores to the fabrics than did the homemaker and student groups. Indications are that the panel was not as pleased with the fabric attributes as were the other groups. Statistical tests were not carried out to measure differences between groups, however.

Summary

A number of research studies involving panel and consumer ratings have been reviewed. A definite pattern of consistency or inconsistency among panel members or between judges and wearers cannot be ascertained. Of the studies in which comparisons of panel members were made, approximately half were in agreement and half in disagreement. Consistency of ratings between panels and consumers was also varied, if mentioned at all. It was noted in several studies that trained panels were more critical than the consumer in their evaluations of textile product performance. Other studies reported close agreement between the evaluations of the two groups.

Although consumer and panel ratings have been used in numerous research studies to evaluate textile product performance, these data have seldom been analyzed statistically to determine rater and consumer consistency. Consequently, it is often not clear how accurate and reproducible such data might be. It is desired that the standards of the consumer be recognized and met by industry in the least expensive manner possible. Understanding consumer needs and variations in the standards of professionals will aid industry and consumers in attaining those ends.

Chapter III

STATEMENT OF THE PROBLEM

One purpose of this study was to determine the consistency of ratings among four professional judges (I, II, III, IV) on four sleepwear fabrics: (1) pink 100% polyester brushed tricot knit, referred to as series 100, (2) lavender 100% nylon tricot knit, referred to as series 200, (3) blue-green flame retardant 80/20% acetate-polyester brushed tricot knit, referred to as series 300 and (4) yellow-green 80/20% acetate-nylon brushed tricot knit, referred to as series 400, in relation to selected performance characteristics. A second purpose of the study was to determine whether the professional panel and the wearer-raters of nightgowns made from the four fabric series given above are consistent in their judgements of selected performance characteristics. The characteristics are as follows:

1. Soil Removal
2. Colorfastness
3. Shape Retention
4. Appearance
5. Hand
6. Durability
7. Static
8. Overall rating

Objectives

The following objectives were developed for this study, using the above-listed characteristics:

1. To obtain ratings of selected nightgown fabrics by wearers and four trained judges on the performance characteristics given above.
2. To determine the consistency of ratings of selected nightgown fabrics between the judges on the performance characteristics given above.
3. To determine the consistency of ratings of selected nightgown fabrics between the trained panel of judges and wearers on the performance characteristics given above.

Hypotheses

Two categories of hypotheses were formulated and used in this study.

Category I Hypothesis: When rating four fabric series (100, 200, 300, 400) relative to the eight performance characteristics (1-8) given above, the level of agreement between a pair of the four professional judges does not exceed the level attributable to chance alone.

Category II Hypothesis: When rating four fabric series (100, 200, 300, 400) relative to the eight performance characteristics (1-8) given above, the level of agreement between one of the four professional judges and the wearer-raters does not exceed the level attributable to chance alone.

Assumptions

For the purposes of this study, the following assumptions were made:

1. The terms describing performance characteristics had identical meaning to all raters.
2. Raters were not affected by garment color, design or fabric composition in rating fabric characteristics.
3. Variation in judgements should not have occurred because of differences in the fabric properties before wear.
4. Each judge rated each garment independently and performance factors were rated independent of one another for each gown.

Limitations

The fiber contents of the sleepwear fabrics used limits the generalization of findings only to those fabrics evaluated. All fabric characteristics may have been evaluated differently by wearer-raters

than by the professional panel. This is expected because performance problems may have been more prevalent during the actual wear testing procedure.

Operational Definitions

Several operational definitions were developed regarding fabric characteristics and these were available to all who evaluated the fabrics.

Fabric characteristics--qualities of the worn fabrics that individuals can evaluate visually and tactually.

1. Soil removal--ease of removing soil and stains
2. Colorfastness -- retention of original color
3. Shape retention--degree of shrinking or stretching in both length and width of nightgown
4. Appearance--changes in fabric such as fuzzing, pilling or snagging
5. Hand--how fabric feels when touched
6. Durability--how well gowns have held up to wear
7. Static--clinging of garments, attraction of lint
8. Overall rating--combined rating of seven fabric characteristics

Chapter IV

PROCEDURE

Selection of Test Fabrics

Fabrics representative of those available for women's nightwear were purchased from one textile producer for use in a previous study (Davis, 1978). Nightgowns were constructed at Virginia Polytechnic Institute and State University under the supervision of an expert in clothing construction. All gowns of each series were cut from one roll of fabric to eliminate variation in fabric lots. The four gown series, all the same style, differed in color and fiber content.

Sample Selection

Fifty-six women consumers were selected for this study because of their availability and willingness to wear and care for nightgowns (Davis, 1978). Another factor considered in wearer selection was that body sizes of wearers were appropriate for the sizes of gowns available. Two gowns were distributed to each wearer in one of six paired combinations. Each woman wore and cared for one of the following sets of gowns: (1) Series 100 and 200, (2) Series 100 and 300, (3) Series 100 and 400, (4) Series 200 and 300, (5) Series 200 and 400 and (6) Series 300 and 400.

Having worn the nightgowns from October 1976 through May 1977, the consumers evaluated the gowns on selected performance characteristics.

The worn gowns were collected and a professional panel evaluated all gowns on the same performance characteristics. This panel was composed of four adult females who are personnel in the Department of Clothing, Textiles and Related Art at Virginia Polytechnic Institute and State University.

Data Collection

Each wearer was given rating cards on which to evaluate the gowns. These cards were selected for use because of their success in a previous study concerning children's sleepwear fabrics (Avery et al., 1978). Each wearer received one card for each gown worn, with the cards color coded to correspond to the gowns. The terms used to describe the performance characteristics were defined on the back of each card to assure clarity (Appendix A).

These cards were modified for use by the professional panel. All gowns of each series were rated on one rating sheet. The characteristics were listed across the top of the sheet and gown numbers recorded vertically. Ratings of excellent, good, fair and poor were recorded numerically as four, three, two and one. This coding was listed at the bottom of each rating sheet to aid the judges. The same definitions given the wearer-raters were made available to the professional judges to minimize variation due to differences in term definitions. Space was provided for the personal reactions and comments of each judge (Appendix B).

Wearer Evaluation

Each consumer wore one set of gowns and laundered them in her customary manner. The majority of gowns were laundered from 25 to 35 times. They rated the gowns on color coded cards on the following performance characteristics:

1. Soil removal
2. Colorfastness
3. Shape retention
4. Appearance
5. Hand
6. Durability
7. Static

Each characteristic was evaluated as excellent, good, fair or poor in comparison to a new 18-inch square sample of the same fabric series. These ratings were assigned the values of four, three, two and one for statistical purposes.

Panel Evaluation

The professional panel members rated all gowns in one room during daylight hours. All gowns of one series were rated on a corktopped 36-inch high table, the surface of which measured 72 by 72 inches. The tables were covered with two layers of white wrapping paper to

insure consistency of background. An adjustable stool was located at the end of each table for the raters' convenience.

The amount and type of light was controlled by drawing the venetian blinds and using fluorescent lighting. Each table was centered under two double fluorescent light panels with the outside edges of the tables and the panels aligned. Each double panel contained eight 48-inch fluorescent tubes which were designated as 40 watt, preheat-rapid start, warm white (F40 WW) and purchased from a major manufacturer.

The unrated, consecutively numbered gowns were stacked in the upper left-hand portion of each table. A control gown, extended in its entirety, was fastened with thumbtacks to the lower left surface of the table. Space to the right of the control was provided to rate the worn garments. Placing the control and worn gowns side by side assured uniformity of light exposure and background. Space for the rated gowns to be placed front down were designated at the upper right section of each table.

A card listing standard ratings for shrinkage was available at each table. Shrinkage of more than one inch received a rating of one, one inch received a two, a half inch received a three and no shrinkage was rated as a four. Rulers were placed at the tables for accurate measurements. The definitions of terms used by the wearer-

raters were also on the tables to aid the judges. Each individual used one rating sheet to evaluate all gowns of one series (Appendix B).

The judges began at different tables and rated all the gowns of one series before moving to the next table. The professionals took a garment from the unevaluated group of gowns and placed it next to the control gown. The worn gown was inspected in comparison to the control on the factors given on the rating sheets. All ratings taken together were used as an overall rating for each gown. The evaluated gown was placed face down in the designated position for the convenience of other raters. This provided the same grouping of consecutively numbered gowns when they were returned to the unevaluated position. All of the evaluations were completed within the span of one week.

Statistical Analysis

Weighted kappa (Cohen, 1968; Cohen, Everitt and Fleiss, 1969) had been proposed as a measure of partial agreement between two judges. No statistic was available to determine agreement among four judges simultaneously. Lin (1975) extended weighted kappa to measure partial agreement among three judges. This statistical technique, even if modified for four judges, does not appear to suit this data. It assigns the same value to perfect agreement between two of three judges, regardless of the rating given by the third judge. For

example, given raters A, B and C, if raters A and B both assign a value of one and rater C assigns a value of two, that case is considered the same as if judges A and B gave a rating of one and judge C assigned a value of four. The above two cases do not constitute the same level of agreement. Therefore, weighted kappa is the most appropriate statistical tool available and was used to determine agreement between the four raters, taken two at a time. The results cannot be pooled to yield a conclusion for all judges simultaneously because the level of significance cannot be determined. Weighted kappa (K_w) is the population parameter of agreement based on the entire population of all possible evaluations. That population parameter is estimated using weighted kappa hat (\hat{K}_w).

The formula for \hat{K}_w follows:

$$\hat{K}_w = \frac{Po' - Pc'}{1 - Pc'} \quad (\text{Cohen, 1968, p. 217})$$

where:

Po' = Proportion of observed agreement

Pc' = Proportion of chance agreement

Chance agreement is calculated from marginal frequencies as in chi square. Observed agreement is computed using weighted values to allow for partial agreement of observations. Regarding the values of these weights, Cohen stated, "These positive weights can be assigned by means of any judgment procedure set up to yield a ratio

scale In many instances, they may be a consensus of a committee of substantive experts, or even, conceivably, the investigator's own judgement The weights assigned are an integral part of how agreement is defined" (1968, p. 215).

A scale of four weights was developed with the aid of a committee of three textile experts and a statistician. A weight of one was chosen for perfect agreement between judges. For a difference of one unit (on a four-point scale) a weight of 0.8 was assigned. This high value was chosen because, in the opinion of the committee, a difference of one did not constitute serious discord. A disagreement of two values was considered more severe and was assigned a weight of 0.3. A range of three points reflected the maximum disagreement and was, therefore, assigned a weight of zero (Table 1).

The null hypotheses were tested using a one tail test with the null hypothesis $H_0: K_W \leq 0$ and alternate hypothesis $H_a: K_W > 0$. A significance level of .005 (critical value = 2.575) was chosen so the test would lead to rejection of H_0 only when \hat{K}_W was considerably larger than zero. Weighted kappa and the z score associated with it were calculated for each pair of judges on each variable. The formula for the z score is: $z = \frac{K_W}{\sigma_{\hat{K}_W}}$ (Cohen, 1968) where $\sigma_{\hat{K}_W}$ is the standard error of weighted kappa, given $K_W = 0$. A more complete explanation of the standard error calculation can be found in Cohen et al. (1969).

Table 1
Weights Assigned to Calculate Proportion
of Observed Agreement

Judge X Ratings	Judge Y Ratings			
	1	2	3	4
1	1.0	0.8	0.3	0.0
2	0.8	1.0	0.8	0.3
3	0.3	0.8	1.0	0.8
4	0.0	0.3	0.8	1.0

Chapter V

FINDINGS AND DISCUSSION

The data for this study were collected as part of a wear study of women's sleepwear. After consumers who participated in the study evaluated their gowns, the garments were collected and rated by a professional panel consisting of four individuals. Upon inspection of the gowns, it was noted that one set of gowns had not been returned. Therefore, data from 27 gowns of series 100 and 300 and 28 gowns of series 200 and 400 were utilized.

The statistical method used to analyze the data is a relatively new technique, which has not been used in the field of clothing and textiles. For this reason, a brief discussion of weighted kappa and the z score associated with it, is warranted. For weighted kappa, a contingency table of the responses of two judges is employed as in a chi-square analysis (Table 2). Preliminary calculations were carried out in which the eight variables were analyzed for each series separately. Examination of the results revealed that a sample, consisting of the observations on a single variable for one series, was too homogeneous to yield useful inferences. The fact that all

Table 2
Contingency Table for Judges I & IV
For Hand of Series 100

Judge I Ratings	Judge IV Ratings			
	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	5	22

gowns were of the same fiber content, together with small sample size, did not allow for a heterogeneous sample. This problem may be best illustrated by an example of a comparison of judges taken from the preliminary data.

Judges I and IV appear to agree when comparing their ratings for fabric hand for series 100 (Table 2). The value of weighted kappa, however, is calculated to be zero. When all observations fall in either one row or column, the proportion of observed agreement (p_0') is equal to the proportion of chance agreement (p_C'). Since $K_W = \frac{p_0' - p_C'}{1 - p_C'}$, when $p_0' = p_C'$, weighted kappa hat will be zero, except when all observations fall into one cell on the main diagonal. In that case, \hat{K}_W is undefined. It can be seen, therefore, that weighted kappa cannot be applied to these data to compare each variable for each series. To obtain useful results using this statistic, the sample must consist of articles of varying quality. A more heterogeneous sample would yield more meaningful results, because variation in the sample would allow for more even distribution of the data in the contingency table. When all ratings fall into a few cells of the matrix, the statistic estimates the judge's overall behavior based on that sample. In such a case, most of the agreement is attributed to chance. A larger sample might also show better results, since more observations are likely to have greater dispersion.

It would be expected that all gowns of one fabric series would perform in approximately the same manner to give a relatively homogeneous sample. Thus, to compare judges' ratings, each variable was considered for all four fabric series combined. This appears appropriate because all four fabrics were designed for the same end use, sleepwear. Comparing all four series at once also gives a larger sample size. The findings are presented in this manner, combining all series for each of eight variables, for statistical purposes.

Weighted kappa hat and z scores were used to test the null hypothesis of $K_W \leq 0$ against the alternative hypothesis of $K_W > 0$. Significance was determined by comparing z scores with the critical value at the .005 level. The small alpha level was chosen so that a value of \hat{K}_W considerably larger than zero is required for rejection of the null hypothesis. Agreement between judges, compared pair-wise, was determined for each of the eight performance characteristics.

Comparisons of Judges

Soil Removal

For the variable of soil removal, all pairs of judges showed significant agreement ($\alpha = .005$). The critical value of 2.575 was exceeded in every comparison and nearly all agreements would be significant at smaller alpha levels (Table 4).

The contingency table of ratings of Judges I and II over all series showed wide dispersion of their ratings on this factor. They seemed to agree that some gowns performed good or excellent, and others fair or poor on soil removal. Regarding each series individually, no distinct pattern of preference could be ascertained. For series 100 and 400, over half of the ratings were either fair or good. Comparisons for series 200 and 300 showed greater variation. As with the overall ratings, the gowns of each of these series were rated from excellent to poor.

The combined series comparisons of Judges I and III did not show a clear consensus as to the performance of the gowns. Series 400 (acetate-nylon) gowns were rated either fair or good 68% of the time. For the other three series, responses were spread throughout the contingency tables, indicating that the judges agreed that while some of the gowns performed satisfactorily, others did not.

Judges I and IV also showed a great deal of dispersion in their comparisons for soil removal over all series. Inspection of the contingency tables for each series separately revealed that, for series 100, 300 and 400, over 50% of the responses given were fair or good, with the remainder for those series spread throughout the tables. No pattern could be determined for the 100% nylon (series 200) gowns.

The judges seemed to be moderately satisfied with the soil removal of series 100, 300 and 400, and agreed that the gowns of series 200 ranged from excellent to poor.

Soil removal was rated good or excellent in 64.5% of the responses over all four series by Judges II and III. This form held true for all series with over half of the responses being good or excellent. This pair of judges felt most of the gowns were highly satisfactory for the factor of soil removal.

The responses of Judges II and IV over all series were more widely varied than were those of the previous pair of judges. Examination of each series separately showed that for series 100 and 400, in over half of the cases, the gowns were rated either fair or good. For series 200 and 300, more than half were rated good or excellent. These two judges appeared to agree that the gowns of all series were generally acceptable.

Approximately 50% of the responses of Judges III and IV, on soil removal over all series, were good or excellent. For each series individually, however, this pattern does not hold true. Over half of the responses for series 100, 200 and 300 were good or excellent, but more than 50% of the gowns of series 400 were rated fair or good by the judges. It may be stated that these judges were generally satisfied with the fabric performance of soil removal of three of the series.

The agreement of judges for the factor of soil removal, was not restricted to one category of responses. There appeared to be variation of fabric performance over all and within each series. While the majority of agreement consisted of ratings from fair to excellent, there were responses at all four levels, from poor to excellent. The variation within each series may be explained by the diversity of care procedures used by the wearers. Some of the women may have been more cautious than others in laundering the gowns.

Colorfastness

The four judges showed agreement, when compared pair-wise, on the factor of colorfastness ($\alpha = .005$). All of the z scores were considerably larger than the critical value and would also be significant at much smaller alpha levels (Table 4).

The ratings of Judges I and II were in nearly every cell of the contingency table when compared for the four series combined. However, for 56% of the comparisons, these judges agreed upon either fair or good evaluations of the gowns. This pattern also held true for the 100% polyester and 100% nylon (series 100 and 200) gowns, with the judges both assigning ratings of fair or good in the majority of cases. Ratings for the acetate-polyester (series 300) gowns were fairly evenly divided among the cells of the contingency table. It would appear that the judges were in accord, but believed that some gowns of that series performed better than did others on colorfastness. For the acetate-

nylon (series 400) gowns, the judges agreed on ratings of poor or fair in 61% of the cases. For 57% of the comparisons, they both felt the gowns were either fair or good. It can be seen that these judges found the 400 series less acceptable than the other series, giving a substantial number of poor or fair ratings.

The agreement between Judges I and III was in the same areas as for Judges I and II. They concurred that a rating of fair or good be given for series 100 and 200 in approximately 50% of the cases. The ratings were spread throughout the table for series 300, showing greater satisfaction with some gowns than with others of that series. For series 400, in 57% of the cases, the judges assigned poor or fair ratings, and in 57% of the cases, the judges agreed on ratings of fair or good. This seemed to indicate less satisfaction with the acetate-nylon (series 400) gowns than with the other three.

In over half of the comparisons, Judges I and IV agreed on ratings of fair or good for all series combined. They also agreed that more than 50% of the 100% polyester (series 100) and the acetate-polyester (series 300) garments showed either fair or good colorfastness. Their ratings of the 100% nylon (series 200) were scattered throughout the contingency table. Poor or fair ratings were assigned to 54% of the acetate-nylon (series 400) gowns and these judges agreed that 61% of these gowns had fair or good colorfastness properties.

Judges II and III agreed that ratings were fair or good in 53% of their comparisons with the four series combined. For the 100% polyester gowns, ratings ranged from fair to excellent. Fifty-six percent of these ratings were fair or good while 63% of the polyester gowns were rated either good or excellent for color. These judges agreed that the majority of the 100% nylon gowns were either fair or good and that over half of the acetate-polyester gowns showed good or excellent color retention. This tends to indicate satisfaction with the gowns of those three series. Less acceptable ratings were given the acetate-nylon gowns, with scores ranging from poor to good. In one-half of the comparisons, ratings of poor or fair were assigned. Ratings of either fair or good were given to 61% of the 400 series garments.

For the variable of colorfastness, Judges II and IV appeared moderately satisfied in the combined series ratings, giving fair or good assessments in the majority of comparisons. Ratings for series 100 and 300 ranged from fair to excellent, with over half of the comparisons yielding ratings of either fair or good and more than 50% good or excellent. Ratings for series 200 were distributed throughout the contingency table, evidence that some gowns performed better than others.

Combined series ratings of Judges III and IV showed agreement of ratings, the majority of which ranged from fair to excellent.

For series 100, ratings of fair or good were assigned in over half of the cases and either good or excellent were agreed upon in 67% of the comparisons. Over half of the responses for series 200 and 300 fell into the good or excellent category, indicating satisfaction with those fabrics. Fair or good ratings were most often agreed upon for series 400, with 71% of the observations in those cells of the contingency table.

A definite pattern of judge response for colorfastness cannot be determined from the above discussion. Perhaps this is due to the wear and laundry procedures used by the consumers. Although ratings varied a great deal, it may be stated that the judges generally found gowns of series 100, 200 and 300 acceptable, but were somewhat less satisfied with the performance of the series 400 garments. These gowns received a greater number of poor ratings than did the other series.

Shape Retention

Agreement on shape retention was found between all pairs of judges ($\alpha = .005$) except for judges I and III (Table 4). For these judges' ratings of shape retention, \hat{K}_W was not found significant at the .005 level. Thus, it cannot be determined if judges I and III agreed in their assessment of that factor. A visual inspection of the contingency table of judges' observations regarding shape retention indicated that

these judges showed major disagreement in over 21% of the comparisons (Table 3). This may have had an effect on the value of \hat{K}_W and the z score used to test the hypothesis.

Judges I and II generally gave low ratings for shape retention over all series. They agreed perfectly that 43% of all gowns were poor on that factor. Sixty-two percent of the ratings were either poor or fair. When each series was examined, it was noted that they were also rated low individually. Over half of all ratings of series 200 and 400 were poor or fair and 93% of the ratings of series 300 were poor or fair. This seemed to show rater dissatisfaction with 200, 300 and 400 series garments. For series 100, the judges appeared more satisfied, giving only 37% of the gowns either poor or fair ratings and 52% of the garments ratings of either fair or good.

The ratings for judges I and IV showed agreement for the series combined, with the majority of garments rated poor or fair. The gowns of series 200 and 300 were also given poor and fair scores in 79% and 85% of the cases, respectively. The ratings of gowns of series 200 and 400 were dispersed throughout the contingency tables. Few scores fell into the excellent category, however. This tends to indicate that Judges I and IV felt some gowns of series 200 and 400 performed satisfactorily for shape retention, while others did not.

In a large majority of the responses (74%), Judges II and III concurred on ratings of poor or fair. This was also true for each

series individually. Poor or fair responses were given to 56% of the 100 series, 79% of the 200 series, 85% of the 300 series and 75% of the 400 series gowns. It appears that these judges found few of the gowns of any series acceptable.

The findings of the comparison of Judges II and IV are similar to the previous pair of judges. In over half the cases for all series combined, and for each series individually, the judges agreed on ratings of poor or fair. In general, they did not seem satisfied with the shape retention of these fabrics.

Judges III and IV also followed the pattern set in the previous two comparisons. More than 60% of all responses were poor or fair for all series combined and separately.

It may be concluded that in most cases, the gowns of all four series were less than acceptable to the judges on the variable of shape retention. In some instances, scores of good were assigned, but seldom were ratings of excellent given. The relatively poor evaluations for the gowns could be attributed to the amount of shrinkage encountered. Judges often commented about shrinkage on the rating sheets.

Appearance

All pairs of judges showed significant agreement at the .005 level on the variable appearance (Table 4). They seemed to agree that

Table 3
Contingency Table for Judges I & III for
Shape Retention for All Series

Judge I Ratings	Judge III Ratings			
	1	2	3	4
1	39.0	21.0	4.0	1.0
2	6.0	6.0	1.0	0.0
3	6.0	7.0	4.0	1.0
4	9.0	3.0	1.0	1.0

while gowns of some fiber contents performed well on this factor, those of other fabrics were less satisfactory.

Judges I and II indicated that 93% of the polyester gowns (series 100) should be rated excellent on appearance. They also agreed that the nylon gowns (series 200) showed exceptionally fine appearance, giving all of those gowns ratings of excellent. These judges appeared much less satisfied with the gowns made of the two fiber blends. Over 90% of all ratings for series 300 were given as poor while 96% of series 400 were rated poor or fair.

The other pairings of judges revealed the same trends in their assessment of gown appearance. Series 100 and 200 were generally found excellent and series 300 and 400 fair or poor. Ratings of excellent were given to 96% of the 100% polyester gowns and to all of the 100% nylon gowns by Judges I and III. By contrast, they assigned values of poor to 81% of the acetate-polyester blends and ratings of poor or fair to 93% of the acetate-nylon blends.

Following the same pattern of agreement, Judges I and IV concurred that over 75% of series 100 gowns were good or excellent and that all of the series 200 gowns had an excellent appearance. Series 300 and 400 gowns were rated much lower, however, with 81% of the acetate-polyester gowns given poor scores and 96% of the acetate-nylon gowns rated poor or fair.

As in the previous pairings, Judges II and III agreed that most of the series 100 gowns (93%) were excellent and all of the series 200 gowns were given that rating. Over 80% of series 300 and 400 were appraised as poor.

Judges II and IV gave a large majority of the 100% polyester gowns an excellent appearance rating. In the opinion of these judges, all of the 100% nylon gowns were excellent, while poor ratings were given to more than 80% of the acetate blended gowns.

Excellent or good ratings were given to 82% of series 100 gowns by Judges III and IV. Like all other pairs of judges, they rated 100% of the series 200 gowns as excellent. Poor or fair scores were given to 89% of the series 300 and to 96% of the series 400 gowns.

From this discussion, it may be concluded that, overall, the judges were extremely pleased with the appearance of the 100% polyester and 100% nylon garments. They seemed to find the gowns of acetate-polyester and acetate-nylon less than acceptable, giving poor ratings to the largest percentage of those gowns.

Hand

A significant level of agreement was shown by all pairs of judges for the factor of hand ($\alpha = .005$) (Table 4). As with the variable appearance, they tended to agree that some fabrics performed well, while others did not.

Hand was assessed as good or excellent for the 100% polyester and 100% nylon garments in over 85% of the ratings by Judges I and II. They found the performance of series 300 and 400 gowns less acceptable, assigning poor or fair in more than 90% of the cases.

Judges I and III agreed completely that the 100 and 200 series gowns were all excellent on the property of hand. They also agreed that the gowns of blended fabrics were not as satisfactory and assigned scores of either poor or fair to all of those garments.

All of the series 100 gowns were given ratings of good or excellent by Judges I and IV. The series 200 gowns were also appraised as good or excellent in 93% of the cases. Similar to the previous pairings, these judges gave all of the series 300 gowns and the majority of the series 400 garments ratings of either poor or fair.

The comparisons of Judges II and III showed that they both felt the 100% polyester and 100% nylon gowns were superior to the acetate blends for the characteristic of hand. Ratings of good or excellent were given for 93% of the series 100 and 89% of the series 200 gowns. By contrast, over 90% of the ratings for series 300 and 400 were poor or fair.

The same pattern of agreement persisted throughout the remaining comparisons. Judges II and IV gave ratings of good or excellent to series 100 and 200 in 96% and 86% of the cases, respectively. Ratings of poor or fair were given by these judges for 96% of the gowns

of series 300 and for 71% of the series 400 garments. Similarly, Judges III and IV agreed that all gowns of series 100 be given good or excellent scores and that series 200 be assigned those ratings in 93% of the cases. All of the acetate-polyester gowns were rated poor or fair by these judges, as were 71% of the series 400 gowns.

Fabric hand is a characteristic that has been considered difficult to assess, depending on the preferences of the rater. These judges tended to show agreement, however, and consistently believed some fabrics were more satisfactory than others. The 100% polyester and 100% nylon gowns were found very acceptable, 85% of them being rated good or excellent. The judges all assigned scores of poor or fair to the majority of the acetate-polyester and acetate-nylon gowns, indicating that the hand of these fabrics was not acceptable to the professional raters.

Durability

All pairs of judges showed significant agreement on the factor of durability across all series at the .005 level (Table 4). The judges were generally more satisfied with the durability of some series than with others.

Judges I and II were satisfied with the durability of garments of series 100 and 200, rating the majority of those gowns good or excellent. Less satisfaction was shown with the 300 and 400 series

gowns. Ratings of poor or fair were assigned to 67% of the acetate-polyester gowns (series 300). The judges did not show a clear consensus on the performance of the series 400 gowns, with scores ranging from poor to good. Eighty-two percent of the gowns were rated fair or good, while 54% were assigned values of poor or fair.

The ratings of Judges I and III varied somewhat from those of the previous pair. While in 85% of the cases they agreed that the series 100 gowns were good or excellent, in half of the cases they felt series 200 gowns were fair or good. The series 300 garments were appraised as poor or fair in 89% of the comparisons. More than 75% of the series 400 garments were rated either fair or good by this pair of raters. These judges seemed the most satisfied with the 100% polyester gowns and the least with the acetate-polyester garments. They appeared to find the 100% nylon and acetate-nylon gowns moderately acceptable for durability.

Judges I and IV agreed that the majority of series 100 and 200 gowns had good or excellent durability. Conversely, they believed that all of the series 300 and 86% of the series 400 garments were poor or fair. Durability of series 100 was rated good or excellent by Judges II and III in 96% of the comparisons. These judges were not as pleased with the series 200 gowns, however, assigning fair or good to over 79% of the garments. Over half of the ratings given

series 300 were poor or fair, indicating a lesser degree of satisfaction with that series. As with the 100% nylon gowns, this pair of judges gave a large majority of the acetate-nylon gowns scores of fair or good.

Judge IV seemed more pleased with the durability of the 100 and 200 series than with the 300 and 400 series gowns. Most ratings given series 100 and 200 gowns were excellent or good. The judges concurred that over half of the 300 and 400 series were poor or fair.

Scores of excellent or good were assigned to 96% of the series 100 (100% polyester) gowns by Judges III and IV. They considered the series 200 (100% nylon) gowns to be fair or good in 71% of the cases. As with all other pairs of judges, this pair felt that most of the series 300 (acetate-polyester) garments (89%) were either poor or fair. No mode of agreement could be ascertained between judges III and IV for the series 400 (acetate-nylon) gowns on durability. In 61% of these comparisons, judge IV gave a score of poor to the garments while Judge II gave a rating of good. Presumably, this amount of disagreement did not affect the value of weighted kappa and the z score enough to show nonsignificant results when all series were combined.

The combined judges' assessments of durability for all series seems to indicate that they agree. A definite pattern of their opinions cannot be determined for all series, however. The judges tended to

agree that the 100% polyester (series 100) gowns performed well, with mostly good or excellent ratings. They generally felt the series 300 gowns were not acceptable, assigning the greatest number of those garments values of poor or fair. The ratings of the other series (200 and 400) showed a less definite pattern, with ratings ranging from fair to excellent for series 200 and poor to good for series 400. One pair of judges lacked close agreement on durability for series 400. These findings might imply that the judges were the most satisfied with the 100 series, followed by the 200 series. The 400 series appeared slightly less desirable than the 200 series garments. The gowns of series 300 seemed the least acceptable to the judges on the variable of durability.

Static

All pairs of judges showed significant agreement ($\alpha = .005$) on the variable of static electricity (Table 4). Some variation was evident in their evaluations of static for each series separately.

Judges I and II gave ratings of fair or good to more than 50% of series 100, 300 and 400, indicating moderate satisfaction with the static properties of these garments. All of the 100% nylon gowns (series 200) were given good or excellent ratings by this pair of judges.

Evaluations of the all-polyester gowns given by Judges I and III ranged from fair to excellent. Over half were assessed as good or

excellent and 59% were rated fair or good. All of the series 200 gowns were given good or excellent scores. The judges showed moderate satisfaction with both the blended fabrics (series 300 and 400), giving scores of fair to good in more than 50% of those cases.

Comparison of Judges I and IV revealed that they viewed the static property of the series 100 gowns as somewhat acceptable, giving ratings of fair or good to 70% of those gowns. As with the previous pairings, all gowns of series 200 were rated good or excellent by these judges. This pair found the acetate-polyester (series 300) gowns less acceptable, assigning values of poor or fair in nearly half the cases. Agreement on the acetate-nylon (series 400) garments was more varied, with ratings ranging from poor to good. Seventy-five percent of those gowns were rated fair or good, while 54% were evaluated as poor or fair for static electricity.

Good or excellent ratings were given the series 100 garments in 70% of the cases by Judges II and III. These judges also appeared pleased with the static performance of the 100% nylon gowns, giving them all good or excellent ratings. More variation was shown in their assessment of the series 300 gowns. Over half of the garments were rated fair or good, while 63% were assigned values of poor or fair. This pair of judges was somewhat more pleased with the series 400 gowns, giving ratings of fair or good to 75% of the garments.

Judges II and IV agreed that fair or good ratings be given the majority of series 100 and 400 garments. They found that the 100% nylon garments had low static levels, giving all those gowns ratings of good or excellent. These judges were less satisfied with the series 300 gowns, giving poor or fair ratings to 67% of the garments.

Ratings of Judges II and IV ranged from fair to excellent for the 100% polyester gowns. They felt that 63% of the gowns were fair or good and 56% were good or excellent. Like all previous comparisons, these judges found all the series 200 gowns to be good or excellent on the property of static. Series 300 garments appeared least acceptable to these judges, with values of poor or fair given to 89% of the acetate-polyester gowns. A large majority (82%) of the acetate-nylon gowns were evaluated as fair or good by these judges with respect to static electricity.

The judges all seemed extremely pleased with the 100% nylon (series 200) gowns on the property of static. The 100% polyester (series 100) gowns appeared to be preferred over the blends. Most of the series 100 ratings ranged from fair to excellent, while ratings of the blended fabrics ranged from poor to good. Series 400 was the more acceptable of the blends, receiving fewer poor ratings than did series 300.

It is appropriate to mention that the judges' ratings were very different from those of the consumers' at the termination of the wear

study. The wearer-raters indicated some dissatisfaction with the static properties of the gowns, overall. This can be attributed to their first hand reactions to high levels of static electricity during wear and care. The professional judges, on the other hand, evaluated the worn garments after they were returned to the laboratory. Therefore, the conditions for evaluation and predispositions of the wearer-raters and judges were dissimilar.

Overall Rating

All pairs of judges showed significant agreement ($\alpha = .005$) when compared for all variables over all series (Table 4). Overall, the judges appeared pleased with the 100 and 200 series garments. They found the acetate blended garments less satisfactory, however. Judges I and II found the 100% nylon gowns acceptable, assigning good or excellent to more than half of those garments. This trend was somewhat evident in the comparisons for each variable separately, with the 100 and 200 series receiving a number of good or excellent scores. The judges were less satisfied with the acetate-polyester and acetate-nylon garments. Over half of these gowns were given ratings of poor or fair, indicating the 300 and 400 series were somewhat unacceptable.

All other pairs of judges showed results similar to those of Judges I and II. In over half of all cases, ratings of good or

excellent were given to the series 100 and 200 gowns. At least 50% of the garments of series 300 and 400 were rated poor or fair, except in one judge pairing. Judges III and IV agreed that 47% of the 400 series were poor or fair. Their remaining ratings were distributed throughout the contingency table.

Overall, it appears that the judges were in agreement concerning the four nightgown series. Although no level of significance can be attached to this agreement, the contingency tables seem to indicate that the judges found that the polyester and nylon gowns performed well and the acetate blended gowns showed a rather low level of overall performance.

Comparisons Between Judges and Wearers

Each judge was compared with the wearers over all four fabric series on the eight performance characteristics. Weighted kappa and the z scores were used to determine agreement at the .005 level of significance for each of the judge-wearer pairs.

Soil Removal

Significant agreement ($\alpha = .005$) could not be determined for any of the judge-wearer pairings for the variable soil removal (Table 5). The consumers appeared more lenient than the judges in most cases.

Table 4
 Weighted Kappa and Z Scores of Agreement Between Judges
 on Performance Factors of All Series Fabrics

Performance Factor	\hat{K}_w	Z Score
<u>Soil Removal</u>		
Judges I & II	0.230	4.557*
Judges I & III	0.116	3.420*
Judges I & IV	0.198	3.307*
Judges II & III	0.349	5.055*
Judges II & IV	0.447	5.724*
Judges III & IV	0.168	2.643*
<u>Colorfastness</u>		
Judges I & II	0.411	6.247*
Judges I & III	0.339	6.169*
Judges I & IV	0.312	5.412*
Judges II & III	0.631	8.272*
Judges II & IV	0.545	6.942*
Judges III & IV	0.453	5.712*

* Significant at .005 level

Table 4
(Continued)

Performance Factor	\hat{K}_w	Z Score
<u>Shape Retention</u>		
Judges I & II	0.246	3.015*
Judges I & III	0.160	2.135
Judges I & IV	0.301	3.665*
Judges II & III	0.442	5.226*
Judges II & IV	0.295	3.344*
Judges III & IV	0.248	2.898*
<u>Appearance</u>		
Judges I & II	0.935	10.241*
Judges I & III	0.901	9.934*
Judges I & IV	0.783	9.048*
Judges II & III	0.936	10.090*
Judges II & IV	0.826	9.357*
Judges III & IV	0.771	8.856*

* Significant at .005 level

Table 4
(Continued)

Performance Factor	\hat{K}_w	Z Score
<u>Hand</u>		
Judges I & II	0.514	7.888*
Judges I & III	0.878	10.001*
Judges I & IV	0.756	9.270*
Judges II & III	0.605	9.102*
Judges II & IV	0.544	8.004*
Judges III & IV	0.708	8.976*
<u>Durability</u>		
Judges I & II	0.502	6.209*
Judges I & III	0.252	3.285*
Judges I & IV	0.510	6.729*
Judges II & III	0.482	6.334*
Judges II & IV	0.463	6.557*
Judges III & IV	0.448	6.048*

* Significant at .005 level

Table 4
(Continued)

Performance Factor	\hat{K}_w	Z Score
<u>Static</u>		
Judges I & II	0.262	3.552*
Judges I & III	0.216	3.222*
Judges I & IV	0.329	4.309*
Judges II & III	0.677	8.575*
Judges II & IV	0.568	7.427*
Judges III & IV	0.590	8.453*
<u>Overall Rating</u>		
Judges I & II	0.521	17.153*
Judges I & III	0.485	16.640*
Judges I & IV	0.521	17.130*
Judges II & III	0.674	22.537*
Judges II & IV	0.591	19.264*
Judges III & IV	0.562	18.681*

* Significant at .005 level

When opinions of Judge I and the wearers were compared over all series, consumers were found to be less critical. In 77% of the cases, the wearers assigned values of good or excellent when poor or fair ratings were given by the judge. This pattern held true when each series was inspected separately. In over 75% of the comparisons for all series (100, 200, 300, 400) the wearers gave higher ratings than did Judge I.

The consumers were less critical than Judge II when compared over all series for soil removal. They gave higher ratings than the judge in 58% of the comparisons. This held for series 100, 300, and 400 also, with the wearers giving higher ratings to more than 60% of the gowns. For series 200, the homemakers were less harsh in their ratings in 43% of the comparisons and harsher than Judge II in only 14% of their ratings. This would seem to indicate that they too were more lenient than this judge.

No distinct pattern could be observed between the comparisons of Judge III and the wearers when compared for soil removal. A number of observations fell into one cell on the diagonal of the contingency table, which may have affected the value of the statistic (see discussion at the beginning of this chapter). The wearers were generally more stringent than the judges in their assessment of the series 100 gowns. They rated them lower 33% of the time and higher in only 15% of the cases. For series 100, 52% of the observations fell into one cell on the diagonal. The wearers were more lenient than Judge III in 43% of the comparisons for series 200, Judge III was more lenient

for only 21% of the 200 series ratings. A number of the observations for the 100% nylon gowns were also concentrated in one cell of the contingency table. A pattern of rater leniency could not be determined for series 300. Sixty-three percent of the observations fell into one diagonal cell of the contingency table.

Over all series for soil removal, Judge IV appeared more critical than the homemakers. In over half of these observations, the judge gave lower ratings than the consumers. When each series was considered separately, the wearers were more lenient in each case. For series 100, 300 and 400, the consumers gave higher ratings in 70% or more of the comparisons. Higher evaluations were given by the wearers for 39% of the 100% nylon (series 200) gowns, while they gave lower scores for only 14% of the observations.

For the factor of soil removal, the wearers seemed more lenient in general than Judges I, II and IV. Judge III and the consumers appeared to agree but \hat{K}_w was not significant. It may be suggested, however, that because many observations fell into one diagonal cell of the matrix, the statistic was influenced.

Colorfastness

When judges and wearers were compared for the factor of colorfastness, no significant agreement was found for any judge-wearer pair at the .005 level (Table 5). In all cases the wearers were less critical than the judges in their evaluation of that factor.

Wearers were considerably more lenient than Judge I in their assessment of colorfastness when all series were combined. In 68% of

Table 5

Weighted Kappa and Z Scores of Agreement Between Judges and Wearers
on Performance Factors of All Series Fabrics

Performance Factor	\hat{K}_w	Z Score*
<u>Soil Removal</u>		
Judge I and Wearers	0.020	1.237
Judge II and Wearers	0.032	0.670
Judge III and Wearers	0.132	2.052
Judge IV and Wearers	0.044	1.122
<u>Colorfastness</u>		
Judge I and Wearers	0.054	1.910
Judge II and Wearers	0.105	2.033
Judge III and Wearers	0.131	2.056
Judge IV and Wearers	0.094	1.530
<u>Shape Retention</u>		
Judge I and Wearers	0.004	0.103
Judge II and Wearers	0.008	0.292
Judge III and Wearers	0.030	1.214
Judge IV and Wearers	-0.013	-0.460

*Critical value = 2.575

Table 5
(Continued)

Performance Factor	\hat{K}_w	Z Score*
<u>Appearance</u>		
Judge I and Wearers	0.085	1.180
Judge II and Wearers	0.118	1.730
Judge III and Wearers	0.115	1.621
Judge IV and Wearers	0.079	1.138
<u>Hand</u>		
Judge I and Wearers	0.052	0.697
Judge II and Wearers	0.040	0.666
Judge III and Wearers	0.093	1.345
Judge IV and Wearers	0.064	0.819
<u>Durability</u>		
Judge I and Wearers	0.040	0.729
Judge II and Wearers	0.078	1.346
Judge III and Wearers	0.101	1.464
Judge IV and Wearers	0.106	2.045

*Critical value = 2.575

Table 5
(Continued)

Performance Factor	\hat{K}_w	Z Score*
<u>Static</u>		
Judge I and Wearers	0.096	1.273
Judge II and Wearers	0.080	1.032
Judge III and Wearers	0.109	1.530
Judge IV and Wearers	0.030	0.377
<u>Overall Rating</u>		
Judge I and Wearers	-0.014	-0.627
Judge II and Wearers	0.054	0.202
Judge III and Wearers	0.096	3.508**
Judge IV and Wearers	0.054	2.204

* critical value = 2.575

** significant at the .005 level

the comparisons, the wearers gave good or excellent ratings when the judge assigned values of poor or fair. Considered separately, each series showed the same results. The wearers gave higher ratings than Judge I in over 80% of the cases.

Comparisons of Judge II and the wearers revealed that the consumers gave higher ratings for colorfastness in 58% of the cases with all series combined. The homemakers were also less severe in their ratings for each series individually. More than 60% of the comparisons for the 100, 200 and 400 series showed the wearers giving higher scores than the judge. For the acetate-polyester gowns, the consumer gave higher ratings than Judge II more than twice as often as they assigned ratings lower than those of the judge.

Ratings of Judge III and the wearers were compared for colorfastness over all series and the wearers were again shown to be less stringent evaluators. The wearers were less critical than the judge nearly three times as often as they were more critical. They gave higher ratings to the 100% polyester gowns in 44% of the ratings while they assigned scores lower than the judge in only 19% of the cases. For the 200 and 400 series garmets, their ratings were higher than the judge's in 50% or more of the observations. Ratings of the acetate-polyester gowns showed less bias, with the wearers only slightly more lenient than Judge III. They gave less severe ratings than the judge for 33% of these garments and were more critical of 26% of the gowns.

In 50% of the observations with all series combined, Judge IV was more critical than the wearers, whereas in only 12% were the consumers harsher than the judge in their assessments of colorfastness. Series 100 and 400 gowns also showed these results with the wearers more lenient in more than 50% of those ratings. The homemakers were also slightly less critical of the 100% nylon garmets, giving higher evaluations than the judge in 32% of the cases and lower values in 25% of the observations. Higher scores were given the series 300 gowns four times as often by the wearers, with 44% higher and 11% of the assessments lower than those of Judge IV.

Colorfastness was generally rated more harshly by the judges than the wearers for all series combined. This trend was seen in every comparison, and in all but two cases the differences were fairly large. This may indicate that the consumers were more satisfied than the judges with the colorfastness of all garments.

Shape Retention

Agreement between the judges and wearers was not significant at the .005 level for any or the comparisons over all series on shape retention (Table 5). Previous discussion revealed that in most instances, the judges were not satisfied with the shape retention of the gowns. The standards of the wearers tended to be somewhat lower than those of the judges on this factor.

Judge I and the wearers appeared to have a difference of opinion as to the shape retention of all garments. The wearers were

consistently less critical than the judge in their assessments of all series combined. In more than 50% of the observations for series 100, 200 and 400, the consumers gave higher ratings than the judge. All of series 300 garments were rated lower by Judge I.

The wearers also showed greater leniency than Judge II in their fabric evaluations regarding fabric shape. In the majority of all cases for each series, the wearers assigned higher scores than did the judges. The same pattern prevailed for the comparisons of Judge III and Judge IV with the wearers. In 60% or more of all cases, the consumers gave good or excellent scores while the judges assigned values of poor or fair for shape retention. Slightly different evaluation techniques were used by the wearers and professional judges. The judge actually measured the shrinkage using an objective scale, whereas, the wearer estimated the amount of change.

Appearance

Appearance was evaluated by both judges and wearers and significant agreement was not found at the .005 level for the judge-wearer pairs (Table 5). When all series garments were considered together, the judges were only slightly more critical than the wearers. Differing patterns of leniency were observed for each series separately, however.

The differences were evident in the comparison of ratings of Judge I and the consumers. The wearers were more severe in their assessments of the 100% polyester (series 100) and 100% nylon (series 200) gowns. In over 50% of these comparisons, lower ratings were

given by the wearers. On the contrary, the consumers gave higher ratings to the blended fabric garments (series 300 and 400) in over 70% of the cases. The findings for Judges II and III and the wearers were consistent with those of the previous pair. Judges II and III were more lenient for the 100 and 200 series garments, assigning higher values than the wearers in the majority of the comparisons. In more than 75% of the observations for series 300 and 400, the wearers gave higher ratings than did these two judges.

Judge IV appeared less lenient than the wearers for the series 100 gowns as well as for the garments of series 300 and 400. In more than half of these comparisons, the judge assigned the lower scores. This judge gave higher values than the homemakers for 75% of the 200 series gowns, however.

A plausible explanation for the differences shown in this discussion is that the term appearance may have been interpreted differently by the judges and wearers. The definition given both consumers and judges was as follows: changes in fabric such as fuzzing, pilling or snagging. Perhaps the wearers considered only those changes mentioned, while the judges may have also considered other changes in fabric appearance.

Hand

Significant agreement was not found between the judge-wearer pairs for the variable of hand ($\alpha = .005$) for all series combined (Table 5). There was a slight tendency for the wearers to be less critical than the judges when this variable was considered with all

nightgown series simultaneously. However, some variation in the standards of the raters was observed for the individual series.

Judge I was less critical than the wearers for series 100 and 200, giving higher ratings of hand to over 60% of those garments. Conversely, the judge rated all of the 300 series garments and 61% of the 400 series gowns lower than did the consumers.

Lower scores were given by Judge II in slightly more of the comparisons made with the wearers for the hand of series 100. In 19% of the cases, the judge was harsher and in 33% the judge was more lenient than the homemakers. The wearers were less critical of the hand of series 200 in 43% of the comparisons and more critical than Judge II in 18% of the cases. The consumers were also more lenient than the judge for the blended fabric series (300 and 400). Seventy-five percent or more of their ratings were higher than the judge's evaluations for those series.

Judge III and the wearers showed differing opinions which followed the same pattern as those for Judge I and the wearers. This judge had higher opinions of fabric hand than the wearers for over 60% of the 100 and 200 series gowns. Judge III had lower opinions of the 300 and 400 series garments, however, giving lower scores than the wearers in over 80% of the comparisons.

As with the previous comparison, the wearers were more critical of the hand of 100 series gowns than was Judge IV. The consumers gave lower ratings in 63% of the cases. The wearers were slightly more critical of the series 200 garments than was Judge IV. The wearers

gave lower ratings than the judge to 36% of the garments, but rated 21% of the gowns higher than did Judge IV. This judge-wearer pair showed the same trend for series 300 and 400 as did all other comparisons for hand. In over 50% of the observations the consumers gave higher ratings than this judge.

Precise trends of judge or wearer leniency cannot be determined for each series for the variable of hand. It can be observed from the preceding discussion that standards of the judges and wearers varied somewhat for the 100% polyester and 100% nylon gowns. In all of the pairings for the acetate blended garments, the wearers appeared to be the less critical graders.

Discussion of judge agreement in a previous section of this chapter revealed distinct patterns of judge agreement on hand for each series. Because hand is a term widely used in the textile field, the judges may have been more familiar with its meaning than were consumers. The tactile sensitivities of individuals varies, perhaps accounting for differences of opinion about the tactile properties of the fabrics.

Durability

A significant level of agreement could not be shown between the wearers and each judge ($\alpha = .005$) for the factor of durability (Table 5). The wearers tended to be more lenient than the judges in assessing fabric durability over all series. Inspection of individual series seemed to support this trend in most of the comparisons.

Judge I appeared to have higher standards than the wearers in the assessment of durability of the 100, 300 and 400 series garments. This judge gave lower scores to more than 50% of those garments. Evaluation of the 100% nylon (series 200) gowns showed the ratings to be about equally divided. The wearers were more stringent than the judge in 36% of the cases and less so in 32% of the comparisons.

Greater leniency was shown by the wearers than by Judge II over all series on durability. This also held true for each series separately. Series 100 garments were evaluated higher by the wearers in 37% of the comparisons and lower only 14% of the time. Over half of the other series garments (200 and 400) were rated higher by wearers than by Judge II.

Comparisons of Judge III and the consumers revealed that over all of the series the homemakers appeared more lenient, giving twice as many higher as lower ratings to all gowns. Consideration of each series individually showed slightly different results, however. For the 100% polyester gowns, the wearers were found to be more severe than the judge in their assessments in 52% of the cases. In the remaining 48% of the comparisons, the judge and wearers were in complete agreement. The consumers gave higher scores than Judge III to more than 50% of both 200 and 300 series garments. They also were somewhat less critical of the acetate-nylon (400 series) gowns, assigning scores higher than the judge to 39% of the gowns and lower values in 21% of the cases.

When all series were combined the wearers seemed to be less critical than Judge IV for durability. This did not hold true for every series when regarded separately, however. Judge IV gave higher scores than the wearers to 48% of the series 100 garments, while the wearers were less critical of only 11% of those garments. The other three series (200, 300 and 400) followed the same trend as over all series combined, with the wearers more lenient in the majority of the comparisons.

Series 100 gowns were given harsher ratings than the wearers by Judges I and II, while Judges III and IV appeared to be somewhat more lenient in their assessments of garment durability. While no definite pattern can be shown for that series, the consumers seemed to find garments of series 200, 300 and 400 more acceptable than did the judges.

Static

Static properties of the four nightgown fabrics were evaluated by the wearers and judges. Significant agreement was not determined at the .005 level for any pair of the judges and consumers over all series combined (Table 5). When the series were combined there was a slight tendency for the judges to be more lenient than wearers for this fabric attribute.

Judge I appeared more critical than the wearers in assessing the series 100 gowns but was less critical of the other series garments. More than three times as many higher as lower values than the wearers

were given by the judge for the 100% nylon (series 200) gowns. Approximately 40% of the 300 and 400 series gowns were rated higher by the judges while the wearers gave higher scores to about 30% of these garments.

Judge II was slightly less critical than the consumers of the static of the 100% polyester gowns. Thirty-three percent of the judge's observations were higher than the wearers' and 22% were lower. This judge was much less stringent than the homemakers in rating the 200 series garments. Nearly 80% of the judge's ratings were higher than those of the consumers. On the contrary, this judge gave harsher ratings to 48% of the 300 series garments and was less critical of those gowns in 30% of the cases. There were apparent differences of opinion between the wearers and Judge II, with them both giving more severe evaluations to about 30% of the 400 series garments.

Static of the 100, 200 and 400 series gowns was assessed more stringently by the wearers than by Judge III in the majority of comparisons. The opposite was true for the acetate-polyester gowns (series 300), with the consumers giving higher ratings nearly twice as often as they assigned values lower than those of the judge.

The wearers were slightly more lenient than Judge IV in their evaluation of the static of the 100 series, giving 41% of the gowns higher scores and 33% lower scores than the judge. As with all other pairings for the 200 series, this judge was less critical of the static than were the wearers. Sixty-four percent of the gowns received

higher ratings from the judge. Conversely, over half of the 300 series acetate blended garments were rated higher by the consumers than by the judge. This judge was only slightly more lenient in the evaluation of the 400 series gowns, assigning higher ratings than the consumers in 46% of the comparisons and lower scores to 36% of the garments.

The wearers and judges appeared to vary in their opinion as to the amount of static in all garments. While the judges were consistently more lenient in evaluating the 100% nylon gowns, reactions to the other fabric series were mixed. It can be seen, however, that leniency was shown more often by the judges than by the consumers. This was an unusual feature of the assessment of static electricity. In comparisons for the other variables, the wearers tended to be more lenient. The consumers probably recalled their experiences with static behavior during the actual wear testing of the gowns. Since the judges did not have actual wear testing experiences to recall, they may have used different standards of evaluation.

Overall Rating

Overall agreement between the judges and wearers was calculated using all responses of the raters over all series. This allowed the researcher to determine judge and consumer agreement for all variables with all series at once. Since there would be more variation in the total sample and a much larger sample size than used for each

variable separately, a more complete measure of judge-wearer agreement could be seen. A total of 770 responses were used in the calculation of weighted kappa in these overall comparisons.

When the judges and wearers were compared overall, significant agreement was not determined for Judge I, Judge II or Judge IV when compared to the wearers. Agreement between Judge III and the wearers was found to be significant at the .005 level, however (Table 5).

Although significant agreement was found between Judge III and the wearers overall, significance was not found when this pair was considered for each of the seven variables individually. When the judge and wearers were compared on soil removal, colorfastness, shape retention, appearance, hand, durability and static, the sample included all four series garments giving a total of 110 observations for each factor. These observations were combined to calculate overall agreement, yielding a much larger sample size. This may explain the significant results obtained in the overall comparison for this judge-wearer pair. The sample for each variable separately may have been too small to yield significant results.

The comparisons of the other judge-wearer pairs showed the homemakers more lenient in each case. Judge I gave lower ratings than the consumers in 59% of the observations and higher ratings in only 19% of all cases. The same trend was evident for Judge II and the wearers, with that judge assigning lower scores than the consumers to 54% of the gowns and higher ratings to 17% of the garments.

Judge IV was also a harsher grader than the wearers, assigning scores lower than the wearers four times as often as higher. In 56% of the cases, this judge gave scores lower than those of the wearers, but gave higher ratings to only 19% of the garments.

When the observations of each judge and the wearers were considered for all seven variables together, an overall measure of agreement was obtained. It appeared that Judge III and the wearers agreed overall. However, since agreement was not determined for any variable separately (i.e., soil removal, colorfastness, etc.), a definite conclusion about the agreement of this judge-wearer pair cannot be made. The wearers were less critical than the other judges when compared overall, giving lenient opinions more often than did those judges.

Chapter VI

SUMMARY

Two main objectives were developed for this study. The first was to determine the consistency of ratings among four professional judges on four selected sleepwear fabrics relative to eight performance characteristics. A second objective was to determine whether the professional panel and the wearer-raters of nightgowns made from the four fabrics were consistent in their judgments of the selected performance characteristics.

The fabrics, which were representative of those available for women's sleepwear included: pink 100% polyester brushed tricot knit, referred to as series 100; lavender 100% nylon tricot knit, referred to as series 200; blue-green flame retardant 80/20% acetate-polyester brushed tricot knit, referred to as series 300; and yellow-green 80/20% acetate-nylon brushed tricot knit, referred to as series 400. Fifty-six women consumers in several locations wore and cared for two gowns of differing fiber content for a winter season. At the completion of the wear period these consumers evaluated the gowns on the following performance characteristics: soil removal, colorfastness, shape retention, appearance, hand, durability and static. An overall rating was determined by analyzing all ratings for the seven factors simultaneously. One woman did not return the two gowns assigned to her. Thus, 110 gowns were individually evaluated in the laboratory by four

clothing and textiles experts. They rated the garments on the same performance characteristics as did the consumers and an overall rating was determined in the same manner as noted above.

Weighted kappa was calculated for each pair of judges and for each judge-wearer pairing. The z scores associated with weighted kappa were also calculated and used to test the hypotheses. A statistic was not available to measure agreement among the judges simultaneously and between the judges as a group and the wearers. Therefore, significance levels reported refer to agreement between judge pairs and between each judge and the wearers. Significance cannot be attached to agreement of the judges or judges and wearers as groups.

Two categories of hypotheses, for all rater comparisons on eight factors were formulated. The researcher rejected or failed to reject the hypotheses on the basis of the findings.

Category I Hypothesis: When rating four fabric series (100, 200, 300, 400) relative to the eight performance characteristics (soil removal, colorfastness, shape retention, appearance, hand, durability, static and overall rating), the level of agreement between a pair of the four professional judges does not exceed the level attributable to chance alone.

A level of significance cannot be determined for all judges simultaneously. Therefore, hypothesis I was tested for all pairs of

judges (I & II, I & III, I & IV, II & III, II & IV, III & IV) for each of the eight variables. Significance levels reported refer to the comparisons between judge pairs.

Agreement between each pair of judges was significant at the .005 level for soil removal. Thus, hypothesis I was rejected for all judge pairs on that factor. The alternate hypothesis of agreement between judge pairs was not rejected. The judges agreed that some gowns performed satisfactorily for soil removal while others were less acceptable.

Hypothesis I was rejected for all judge pairs because agreement was found significant for each pair of judges in rating colorfastness ($\alpha = .005$). The alternate hypothesis of judge pair agreement was not rejected. As with the factor of soil removal, the judges agreed that some of the nightgowns performed better than others for colorfastness.

All judge pairs were compared for the factor of shape retention. The first hypothesis was not rejected ($\alpha = .005$) for all judge pairs except Judges I and III. The critical value of the statistic was not exceeded for that judge pair. Thus, hypothesis I was not rejected for Judges I and III but was rejected for the remaining judge pairs. When agreement between the judge pairs was supported, a degree of dissatisfaction with the shape retention of the garments was noted. This may have been due to the fabric shrinkage observed by the raters.

Agreement was found to be significant at the .005 level for each judge pair on the factor of appearance. Therefore, hypothesis I was rejected for the judge pairs and the alternative hypothesis of judge agreement was accepted for each pair. Generally, the judges seemed to find the 100% polyester and 100% nylon gowns (series 100 and 200) acceptable and were less satisfied with the acetate blended garments (series 300 and 400).

Hypothesis I was rejected for all judge pairs for the variable of hand ($\alpha = .005$). The alternate hypothesis of judge agreement was accepted for the judge pairs. As with the factor of appearance, the judges seemed more satisfied with the performance of the 100% polyester and 100% nylon gowns. They also found the hand of the acetate-polyester and the acetate-nylon gowns generally less acceptable.

Durability of the nightgowns was rated by the judges. Agreement was found to be statistically significant ($\alpha = .005$) for all judge pair comparisons for that factor. Therefore hypothesis I was rejected for each of the judge pairs and the alternate hypothesis of agreement was accepted for any rater pair. The judges generally seemed to prefer the 100% polyester (series 100) garments, followed by the 100% nylon (series 200) gowns. They appeared to view the acetate-nylon (series 400) gowns as less acceptable and were the least satisfied with the acetate-polyester (series 300) gowns.

Agreement between each pair of judges was found significant at the .005 level for the variable of static. Hypothesis I was rejected

for all judge pairs and the alternate hypothesis was accepted for that factor. The judges seemed to agree that the 100% nylon gowns had the least static problems, followed by the 100% polyester gowns. They appeared less pleased with the performance of the acetate blended garments, with the acetate-polyester gowns displaying the worst static problems.

When the above seven variables (soil removal, colorfastness, etc.) were combined for analysis, an overall measure of agreement was obtained. The first hypothesis was rejected for all judge pairs when compared for the overall ratings. Agreement was determined significant for all judge pairs at the .005 level. Hence, the alternative hypothesis of judge agreement was accepted for each judge pair. Overall, it was noted that the judges were most satisfied with the performance of the 100% polyester and 100% nylon garments and found the acetate blended gowns generally less acceptable.

A second category of hypotheses was used, comparing each judge with the wearers as a group. As with the first category of hypotheses, the significance levels reported pertain to each judge-wearer pairing.

Category II Hypothesis: When rating four fabric series (100, 200, 300, 400) relative to the eight performance characteristics (soil removal, colorfastness, shape retention, appearance, hand, durability, static

and overall rating), the level of agreement between one of the four professional judges and the wearer-raters does not exceed the level attributable to chance alone.

Hypothesis II was not rejected for the variable of soil removal for each judge when compared to the wearers ($\alpha = .005$). In the majority of the comparisons, the consumers appeared to be more lenient than the judges in evaluating the gowns.

Agreement ($\alpha = .005$) was not significant between any of the judges and wearers for the variable of colorfastness. Therefore, hypothesis II was not rejected for all judge-wearer pairs. It was observed that the consumers were generally more satisfied than the judges with the colorfastness of the gowns.

Hypothesis II was not rejected for any judge-wearer pair on the factor of shape retention. Agreements could not be shown significant for any pairing at the .005 level. The wearers appeared less critical than the judges in their assessments of shape retention of the garments.

Comparisons of judge and wearer evaluations for the variable appearance showed that agreement was significant ($\alpha = .005$) for any judge-wearer pair. Thus hypothesis II was not rejected for any judge-wearer pair. A pattern of judge or wearer leniency could not be determined for the factor of appearance. It is possible that the term appearance had different meanings to the wearers and judges, even though

identical definitions were available to both groups. This could account for the lack of a trend in the standards of the raters.

Agreement between each judge and the wearers was not found significant at the .005 level for the factor of hand. Therefore, the second hypothesis was not rejected for each judge-wearer pairing. As with the variable of appearance, no trend in the standards of the wearers or judges could be determined. A possible explanation is that hand is a term which is more familiar to clothing and textiles professionals than to the consumer. The evaluations of this property may also depend largely on the personal preferences of the individual raters.

Hypothesis II was not rejected for each judge-wearer pair for the characteristic of durability. Agreement could not be found significant for any of the pairings at the .005 level. In the majority of the comparisons, the consumers tended to be more lenient than the judges in their evaluations of fabric durability.

Static was evaluated by the judges and wearers of the gowns. The second hypothesis was not rejected for all judge-wearer pairs for that factor because agreement was not shown significant for each pair at the .005 level. While a pattern of rater leniency was not extremely evident, there was a slight tendency for the wearers to be more critical than the judges in their assessment of this factor. Because static properties may have been more evident during the wear testing period, the consumers were probably exposed to the static problem to a greater degree than were the judges.

To determine an overall measure of agreement, the observations of the judges and wearers were combined for all variables (soil removal, colorfastness, etc.) for statistical analysis. Hypothesis II was not rejected for Judge I, Judge II and Judge IV when each was compared to the wearers on the overall rating. Agreement was not significant at the .005 level for the above judge-wearer pairs. Agreement between Judge III and the wearers was found to be significant at the .005 level, however. The second hypothesis was rejected for this pairing for the overall comparisons. It must be noted, however, that agreement was not found significant when Judge III was compared with the wearers for each variable separately (soil removal, etc.). Therefore, a definite conclusion about the agreement of this judge and the wearers cannot be stated. Overall, the other judges appeared more critical than the wearers, giving harsher ratings to the garments in most cases.

Recommendations for Future Research

The need for further research became evident during the course of this investigation. The development of a statistical procedure to determine agreement among a number of raters simultaneously is important to the continuation of research in this area. This would not only simplify the statistical analysis but would also allow for more concrete conclusions based on appropriate significance levels. The weighted values used to determine partial agreement should also be evaluated to determine the validity of the weights that were employed.

A more heterogeneous sample than that used in this study would lend itself better to statistical analysis. This could be accomplished by employing fabrics of differing strengths, abrasion resistance and other properties which would cause them to perform differently under similar wear conditions. If ratings among members of a professional panel only were to be compared, fabrics could be subjected to various laboratory treatments to simulate different levels of performance.

The factor of overall rating should be added to the satisfaction record cards and judges' ratings sheets in future studies. This would indicate the raters' total satisfaction with the garments.

This investigation should be expanded to include other groups of professional raters and consumers. An attempt should be made to include consumers who are representative of the total population. Comparisons of ratings of fabrics suitable for end uses other than sleepwear could also be undertaken.

REFERENCES

- AATCC Technical Manual. Research Triangle Park, N.C.: American Association of Textile Chemists and Colorists, 1977.
- Avery, C., Schaeffer, K., Densmore, B., Ahren, V., Mauldin, W., Peters, E., Quick, E., Smith, B. and Zentner, M. A. Consumer, market and laboratory studies of flame resistant textile items, part III: pajama wear study. University of Rhode Island Experiment Station Bulletin 424, June 1978.
- Ball, H. M. The appearance of men's worn and nonworn durable press work trousers. Unpublished doctoral dissertation, Texas Women's University, 1971.
- Bercaw, J. R. New developments in evaluating "wash-and-wear" performance. American Dyestuff Reporter, 1960, 49 (4), 37-43.
- Booker, C. A. A comparison of a stretch woven fabric with a conventional woven fabric by laboratory and wear testing. Unpublished master's thesis, The University of Tennessee, 1974.
- Bostrum, A. J. Appearance of sheets composed of cotton and blends of cotton with regular and high modulus polyester. Unpublished doctoral dissertation, Texas Women's University, 1973.
- Bradley, V. R. Performance of in-service consumer constructed durable press shirts. Unpublished master's thesis, Southern Illinois University, 1972.
- Chamberlain, L. K. A consumer evaluation of men's and women's Qiana knit sportshirts. Unpublished master's thesis, Southern Illinois University, 1972.
- Christian, D. S. A comparative study of the effects of drycleaning and laundering on the appearance and performance of a selected foam knit fabric. Unpublished master's thesis, University of Alabama, 1972.
- Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70 (4), 213-220.
- Cohen, J., Everitt, B. S., Fleiss, J. L. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 1969, 72 (5), 323-327.

- Davis, J. Consumer satisfaction, preferences and care procedures for selected women's sleepwear fabrics. Unpublished master's thesis, Virginia Polytechnic Institute and State University, 1978.
- Donnel, M. J. Construction, care and performance of in-service deep pile fabric overshirts. Unpublished master's thesis, Illinois State University, 1965.
- Erickson, N. C. The comparative evaluation of performance of 65/35 polyester/cotton blend shirts with permanent press and soil release finishes when subjected to wear and drying treatments. Unpublished master's thesis, University of Arizona, 1971.
- Fargas, D. E. H. Consumer evaluation of eighty-two percent nylon/ eighteen percent spandex stretch lace dresses. Unpublished master's thesis, Southern Illinois University, 1972.
- Freeman, M. R. Effectiveness of commercial applied and consumer applied soil release finishes on white polyester-cotton durable press fabric. Unpublished master's thesis, Southern Illinois University, 1972.
- Harper, P. B. An investigation of pre-cured permanent press fabric for the home sewer. Unpublished master's thesis, The University of Tennessee, 1968.
- Headrick, B. M. Laboratory and end use performance of white table napkins made from a soil releasing durable press fabric compared with a durable press fabric. Unpublished master's thesis, The University of Tennessee, 1968.
- Inlow, J. L. Serviceability and consumer appeal of the plain weave, terry weaver and stitch-through kitchen towels. Unpublished master's thesis, Southern Illinois University, 1966.
- Larson, E. A. S. Comparative evaluations of men's durable press trouser fabrics. Unpublished master's thesis, Texas Women's University, 1973.
- Laughlin, K. C. Consumer acceptance of FR sleepwear. Textile Chemist and Colorist, 1976, 8 (3), 51-55.
- Laughlin, K. C. and Buddin, F. Consumer acceptance of FR sleepwear: Part II. Textile Chemist and Colorist, 1977, 9 (9), 70-74.
- Lin, Y. S. Weighted kappa for three raters. Reprint from Social Statistics Section, Proceedings of the American Statistical Association, 1975, 529-531.

- Morris, M. A., Margerum, B. J. and Young, M. A. Performance of durable press shirts. Textile Chemist and Colorist, 1972, 4 (7), 39-42.
- Murdock, J. E., Jr. How the army tests fabric wear. Textile Chemist and Colorist, 1969, 1 (23), 18-21.
- Ogle, E. M. An analysis of the performance of durable press shirts: Part II, worn and commercially laundered twenty-five times. Unpublished master's thesis, The University of Tennessee, 1968.
- Paek, S. Response profiles of flame-retardant fabric hand. Textile Research Journal, 1978, 48 (8), 487-489.
- Perenich, T. A. Effects of light exposure and dark storage on polyester-cotton blend fabrics with optical brighteners. Unpublished master's thesis, The Florida State University, 1972.
- Powderly, D. D. The changing standards of the consumer. Modern Textiles, March 1970, 51, 88-91.
- Ruppenicker, G. G., Rhodes, A. R., Markezich, A. R. & Little, H. W. Crease damage ratings are useful for evaluating fabric edge abrasion. Textile Chemist and Colorist, 1972, 4 (6), 33-35.
- Suansamroeng, S. General appearance, dimensional stability and durability of men's shirts after wear and laundering. Unpublished doctoral dissertation, Texas Women's University, 1974.
- Tyler, L. M. A study of seaming and serviceability properties and their retention in polyester-cotton blends of varying percentages. Unpublished master's thesis, Cornell University, 1972.
- Whitcomb, M. G. The wear testing of fabrics. Modern Textiles, March 1969, 50, 64-67.
- Wright, E. L. & Clary, F. L. Design and operation of wear test experiments. Proceedings of the Textile Wear Test Symposium, 1969.
- Yuan, L. S. Practicality of selected disposable laboratory coats as evaluated by wear and laboratory tests. Unpublished master's thesis, Southern Illinois University, 1967.

APPENDICES

APPENDIX A

WEARER RATING CHECK SHEET

Satisfaction Record
Date _____

Name _____

	Excellent	Good	Fair	Poor
1. Soil removal				
2. Colorfastness				
3. Shape retention				
4. Appearance				
5. Hand				
6. Overall durability				
7. Static				
8. Personal reactions and comments				

EXPLANATION OF TERMS

1. Soil removal--ease of removing soil and stains.
2. Colorfastness--retention of original color.
3. Shape retention--degree of shrinking or stretching in both length and width of nightgown.
4. Appearance--changes in fabric such as fuzzing, pilling, or snagging.
5. Hand--how fabric feels when touched.
6. Overall durability--how well gowns have held up to wear.
7. Static--clinging to garments, attraction of lint.
8. Personal reactions and comments--statements to explain the above responses to the rating factors, comments to describe nightgown characteristics not covered by the above rating factors.

APPENDIX B

PROFESSIONAL RATER CHECK SHEET

Prof. Rater _____

NE - 79 Nightgown Wear Study

June 1977

Gown No.	Soil Removal	Color-fastness	Shape Retention	Appearance	Hand	Durability	Static	Personal Reactions & Comments

4 = excellent; 3 = good; 2 = fair; 1 = poor.

Definitions of Terms

1. Soil removal--ease of removing soil and stains.
2. Colorfastness--retention of original color.
3. Shape retention--degree of shrinking or stretching in both length and width of nightgown.
4. Appearance--changes in fabric such as fuzzing, pilling and snagging.
5. Hand--how fabric feels when touched.
6. Durability--how well gowns have held up to wear.
7. Static--clinging to garments, attraction of lint.
8. Personal reactions and comments--statements to explain the above responses to the rating factors, comments to describe nightgown : characteristics not covered by the above rating factors.

**The vita has been removed from
the scanned document**

CONSISTENCY OF JUDGMENTS OF A PROFESSIONAL PANEL AND WEARERS
REGARDING CHARACTERISTICS OF SELECTED SLEEPWEAR FABRICS

by

Margaret Jane Gallagher

(ABSTRACT)

The purposes of this study were to determine the consistency of subjective ratings among four professional judges and between those judges and wearers of four fabric series nightgowns. The fabrics included: (1) 100% polyester brushed knit; (2) 100% nylon knit; (3) 80/20% acetate-polyester brushed knit and (4) 80/20% acetate-nylon brushed knit. Data were collected from 55 women who each wore and cared for two garments of differing fiber contents and evaluated them on eight performance characteristics.

The consumers rated the fabrics on the following characteristics: soil removal, colorfastness, shape retention, appearance, hand, durability and static. An overall rating was obtained by combining data for the above seven factors. The gowns were also evaluated by four professional raters on the above factors. An overall rating was also determined for these judges.

Weighted kappa and z scores were used to measure agreement between judge pairs and between each judge and the wearers for the eight factors. The researcher accepted the hypothesis that rater agreement exceed that expected by chance for all judge pairs on all factors except

shape retention. However, for one judge-pair the hypothesis of agreement was not accepted for shape retention. The hypothesis of rater agreement was rejected for each of the judges and the wearers on all factors except overall rating. The hypothesis of agreement was accepted for Judge III and the wearers for the overall rating. The consumers were generally more lenient than the professionals in their evaluations of all fabric characteristics except static electricity. The judges were less critical than the wearers in assessing that factor.