UNIVERSITY OF ALBERTA

IMPACT OF RATERS' LEVELS OF RESPIRATORY TRAINING AND EXPERIENCE ON THE QUALITY OF SPIROMETRIC INTERPRETATIONS IN EPIDEMIOLOGICAL STUDIES

by

TANIA STAFINSKI ©

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

in

MEDICAL SCIENCES - PUBLIC HEALTH SCIENCES

EDMONTON, ALBERTA
FALL, 2000

0-612-59881-0

University of Alberta

Library Release Form

Name of Author: Tania Stafinski

Title of Thesis: Impact of Raters' Levels of Respiratory Training and Experience on the Quality of Spirometric Interpretations In Epidemiological Studies

Degree: Master of Science

Year this Degree Granted: 2000

Tania Stafinski

Tania Stafinski, BSc
20 Varston Place N.W.
Calgary, AB
T3A 0B7

May 26, 2000
Date

ABSTRACT

Spirometry is a fundamental part of respiratory epidemiology. Its use involves subjective interpretation of graphic output. This study investigated the type of respiratory expertise required to properly assess spirogram acceptability. Two spirogram sets, 1 comprising tracings generated for this study and 1 incorporating curves from previous epidemiologic research, were interpreted by 4 categories of raters whose respiratory expertise levels varied. Within and between-category concordance and the effect of spirometry technician expertise on inter-rater reliability were assessed using kappa. Principal Components Analysis applied to kappa matrices identified patterns of rater agreement. Logistic regression examined participant characteristics associated with test acceptability. Results indicated no relationship between rater agreement and level of respiratory expertise. Greater concordance was not definitively correlated with spirograms from highly trained technicians. No demographic or cardiopulmonary health-related characteristic of participants contributed to test acceptability. Lastly, previous spirometry exposure did not increase the likelihood of performing an acceptable test.

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *Impact of Raters' Levels of Training and Experience on the Quality of Spirometric Interpretations in Epidemiological Studies* submitted by Tania D. Stafinski in partial fulfillment of the requirements for the degree of Master of Science in Medical Sciences-Public Health Sciences

Patrick A. Hessel, PhD

Richard L. Jones, PhD

Angus H. Thompson, PhD

Donald Schopflocher, PhD

Eric Y. L. Wong, MD, MSc

23-05-2000

Date of Thesis Approval by Committee Members

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to the following individuals for their invaluable assistance with and contributions to this study.

From inception to completion, my supervisor, Dr. Patrick Hessel, provided continuous inspiration, support, encouragement, and advice. I am privileged to have learned the principles of epidemiology from a truly exceptional teacher.

My committee members offered conscientious, insightful expertise. Dr. Richard Jones was instrumental to the study's development and design. Dr. Don Schopflocher and Dr. Angus Thompson generously shared their statistical ingenuity and direction during the analysis phase.

Thank you to Dr. Eric Wong for his thorough editorial review.

The understanding and confidence offered by my friends throughout this project was especially appreciated. In particular, I am grateful to those who critically appraised my thesis drafts.

Lastly, I wish to thank my parents who not only fostered a desire to set academic goals but also supplied me with the tools required to realize them.

TABLE OF CONTENTS

CHAPTER THREE: RESULTS - CHARACTERISTICS OF THE STUDY POPULATIONS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION AND BACKGROUND

This chapter establishes spirometry as a principle means of evaluating respiratory health. Further, it acknowledges concerns regarding the reliability of spirometric data collection and interpretation, two factors which directly influence spirometry's clinical and epidemiologic utility.

1.1 Definition of Spirometry

Pulmonary-function tests generate objective, quantifiable assessments of respiratory status (Hughes and Empey, 1981). The most basic and widely used such test is the spirometric examination (Bosse, 1993). Spirometry measures the volume of air inhaled and exhaled from a subject's lungs as a function of time during simple, clearly defined, breathing maneuvers (Crapo, 1994). Subsequent critical inspection of the graphic records or spirograms produced can indicate changes in functional condition and disease state (Wanger, 1992).

1.2 Role of Spirometry

Since "spirometric results correlate well with morbidity and life-expectancy" they provide essential, physiologic evaluations throughout the diagnosis, treatment, and monitoring phases of pulmonary clinical case management (Table 1.1) (ATS, 1995).

Table 1.1 Clinical Indications for Spirometry*

Diagnostic
To evaluate symptoms, signs, or abnormal laboratory tests
To measure the effect or severity of disease on pulmonary function
To detect sub-clinical abnormalities in high-risk patients (e.g., heavy smokers)
To assess pre-operative risk in patients undergoing thoracic or upper-abdominal surgery
Monitoring (serial measurements)
To determine the effectiveness of therapeutic interventions
To trace the course of diseases affecting lung function
To evaluate the current status of individuals with occupational exposure to injurious substances
Disability/Impairment Evaluations
To assess patients as part of rehabilitation programs
To evaluate individuals for insurance or legal reasons

*Adapted from Crapo, 1994.

Spirometric testing is not restricted to a clinical (i.e., patient) context. The heightened awareness of chronic lung disease as a major public health concern continues to intensify the need for effective, pulmonary surveillance strategies (Hankinson, 1986). Such strategies require simple, non-invasive, economical tests that feature adequate sensitivity[1] and specificity[2]. Spirometry addresses and, more importantly, satisfies each of these criteria (Petty, 1997). Therefore, beyond its conventional clinical use, it has become a fundamental part of routine medical screening. Complementary to this role is its employment in population-based research (Becklake et al, 1993). During the past four decades, studies conducted by the United States' National Heart, Lung, and Blood Institute generated 130 000 sets of spirometric results representing more than 50 000 participants. Collectively, they have produced a "wealth"of information available for collaborative, scientific investigation (Manolio, 1997). Thus, because of its capacity to serve as a valuable indicator of general respiratory health, spirometry is recognized as an essential tool in pulmonary epidemiology (Table 1.2) (Ruppel, 1997).

Table 1.2 Public Health Indications for Spirometry*

Public Health Screening
Screening populations to detect pulmonary disease in individuals who are not yet receiving medical attention
Epidemiologic Research
Comparison of the health status of populations in various environments
Validation of subjective complaints in occupational and environmental settings
Derivation of reference equations

*Adapted from Crapo, 1994

1.3 Spirometry in Epidemiology

Respiratory epidemiologic investigations often involve field administration of spirometry. Data are assembled in various environments where access to traditional

[1] Sensitivity is the proportion of truly diseased persons in a screened population who are identified as diseased by the screening test. Thus, it is the probability of correctly diagnosing a case (Last, 1995).

[2] Specificity is the proportion of truly non-diseased persons who are identified as such by the screening test. Thus, it is the probability of correctly identifying a non-diseased person with a screening test (Last, 1995)

research facilities (i.e., pulmonary function laboratories) with stringent control is limited (Ferris, 1978). Consequently, test quality remains a significant concern (Crapo, 1994). Epidemiologic studies, in particular, are characterized by large quantities of reports requiring not only accurate, but, also, efficient interpretation (Rothman, 1996). Therefore, in order to maximize data quality, minimize bias inherent in multi-centered surveys, and, ultimately, ensure comparability of results, researchers, adopting spirometry as part of their methodology, devised and compiled a comprehensive set of standards promoting accurate and appropriate spirometric data collection. The Epidemiology Standardization Project, published in 1978, provided equipment recommendations, procedural guidelines, and a detailed test interpretation strategy (Ferris, 1978). As a supplement to this document, the American Thoracic Society (ATS) developed a coinciding set of criteria in 1979. From inception, the ATS, with the support of the European Respiratory Society, has continually reviewed and revised these criteria to reflect concerns predominately emerging from epidemiologic research (ATS, 1995).

1.4 Spirometric Test Quality-Acceptability and Reproducibility Criteria

Although encouraged but not mandated, most studies employing spirometry, irrespective of test environment (i.e., field, occupational, or clinical), adhere to ATS protocol, thereby ensuring optimal quality (Becklake et al, 1993). Defined within this protocol are criteria for evaluating spirometric test acceptability. Application of these criteria to each test's graphic record determines its quality and, in turn, its validity (Table 1.3).

Table 1.3 Summary of Quantitative and Qualitative Acceptability Criteria*

Satisfactory start of exhalation
Evidence of maximal effort
No hesitation-extrapolated volume less than 5% of FVC or 0.15 L, whichever is greater
No false start
No cough or glottis closure, especially during the first second of exhalation
Satisfactory duration of test
At least 6 seconds and/or a plateau in the volume-time curve
Reasonable duration or a plateau in the volume-time curve - no change in volume for at least 2 seconds
No evidence of mouthpiece obstruction
No evidence of a leak

*Adapted from American Thoracic Society, 1994

Additionally, spirometric examinations comprise multiple trials (i.e., maneuvers) per subject test session. Analysis of the generated tracings includes identification of "reproducible" or consistent curves from those already judged as acceptable (Wise et al, 1995). ATS-established reproducibility criteria require that the curves' forced vital capacity (FVC)[1] and forced expiratory volume (FEV_1)[2] values, both calculated at the time of the test, be within 5% or 200 ml of each other (whichever is greater). Tests not meeting both ATS acceptability and reproducibility standards are defined as "test failures" (ATS, 1995). A prerequisite to the application of reproducibility guidelines is initial compliance with acceptability guidelines. While the ATS suggests that a test not be rejected solely on the basis of its lack of reproducibility, it must be rejected when acceptability criteria are not satisfied. Thus, acceptability serves as the fundamental determinant of test quality (ATS, 1991).

Because these curves constitute the data set for all subsequent evaluations, it is imperative that the selection be correct (Behringer, 1991).

1.4.1 Qualitative and Quantitative Guidelines for Assessing Acceptability

Compliance with acceptability standards requires consideration of both quantitative and qualitative guidelines (Ruppel, 1997).

Quantitative criteria refer to computable parameters. For example, using a back extrapolation[3] technique, time zero is determined and forced expiratory volumes (FEV_t) are calculated (Laszlo and Sudlow, 1983). ATS-acceptable maneuvers should yield an extrapolated volume of less than 5% of the forced vital capacity (FVC) or 0.15 L

[1] FVC (forced vital capacity): the maximal amount of air forcefully exhaled from the subject's lungs after full inspiration (Wanger, 1992)

[2] FEV_1 (forced expiratory volume): the volume of gas forcefully expired during the first second of an FVC maneuver (Wanger, 1992)

[3] Back extrapolated volume refers to the volume of gas that escapes into the spirometer before the subject achieves maximal flow (Wanger, 1992).

(whichever is greater). Secondly, a minimum 6 second exhalation time and/or maintenance of a 1 second plateau, exhibiting "no change in volume", is required (Table 1.3) (ATS, 1995). These numerical recommendations for both 'start of test' and 'end of test' assessment protect against excessive hesitation and premature termination, respectively (Miller, 1987). Most automated spirometers are programmed with ATS-quantitative criteria-derived computer algorithms into which the calculated parameters from collected tracings are incorporated (Clausen, 1986). The resulting report includes a statement indicating the test's acceptability. Despite documented reliance on these numerical statements, ATS studies reviewing actual curves confirmed that unacceptable data can be obtained from tests meeting quantitative criteria, exclusively (Kunzli, 1995). Thus, critical inspection of the graphic records following qualitative guidelines becomes necessary (Ruppel, 1997).

Qualitative criteria require a subjective evaluation of the spirograms (McKay et al, 1991). Each flow-volume or volume-time curve is visually appraised, noting whether or not the test subject achieved full inspiration prior to beginning the FVC maneuver and exhibited maximal effort throughout the maneuver, without hesitation, coughing, or glottis closure (Glindmeyer, 1987). A smooth, uninterrupted contour demonstrates freedom from such artifacts (Table 1.3). Tests failing to comply with any of these restrictions suggest unacceptable performance quality and, in turn, are rejected (Chusid, 1983).

1.4.2 Stages of Determining Acceptability

There are two stages at which test acceptability can be determined: 1) at the time of test administration or 2) at the point of test interpretation (Chusid,1983). During the spirometric procedure, subsequent to conducting the effort-dependent maneuver, a technician inspects the results. This provides the initial opportunity for judging test acceptability (Wanger, 1992). Results, then forwarded to an interpreter, receive a second analysis (Quadrelli et al, 1996). Regardless of stage, interpretations utilize the same set of qualitative and quantitative acceptability criteria (Hughes and Empey, 1981). As previously stated, qualitative guidelines contribute subjective evaluations whereas those quantitatively-based deliver objective, numerical appraisals. Although the importance of both have been well-established, the qualitative guidelines allow for greater variance in the assessment of test acceptability (Ruppel, 1997). This variance facilitates the measurement of agreement/disagreement among interpreters (inter-rater reliability) at either of the two stages (Last, 1995). Further, the level of training and experience of the interpreters may vary, potentially influencing their judgement (Wise et al, 1995).

1.5 Spirometric Test Acceptability in Relation to Pulmonary Expertise

Past studies have examined each interpretation stage in their attempt to clarify the relationship between pulmonary expertise (i.e., training and /or experience) and assessment of spirometric acceptability (Enright, 1991).

1.5.1 Stage 1: Test Administration

In 1991, the American Thoracic Society's statement on lung function testing proclaimed that the largest single source of error resulting in "test failure (was) improper performance of the test, itself"(ATS, 1991). Current research also supports this conclusion (Ruppel, 1997). Thus, the technician's roles as both the test's administrator and its interpreter (i.e., of acceptability) become critical. By carefully observing the subject and the corresponding curves, he/she "must elicit vigorous subject effort, be able to recognize faulty technique", and advise corrective action (Wenzel and Larsen, 1996). The perceived expertise required for maximizing test quality is defined by the protocol of the individual research investigation (Becklake, 1993).

One recent study, designed specifically to investigate spirometric test acceptability in primary care practices, determined the impact of training on spirometry performance. Although not their principal objective, five additional epidemiologic studies also examined the relationship between reliable interpretations of spirometric acceptability and technician expertise (i.e., training and experience).

i) Study of Technician Expertise Level and Test Acceptability

New Zealand Primary Care Practice Study

In 1998, a New Zealand-based, prospective, intervention study examined the influence of "formal training" on the quality of spirometry in a clinical setting (Eaton et al, 1999). To initiate the16-week investigation, 30 voluntary primary care practices, each contributing one doctor and one nurse, were randomly selected and arbitrarily divided into two categories: "trained" and "usual". Those defined as "trained" attended two spirometry administration workshops: an introductory program at week 0 and a maintenance session at week 12. In contrast, those defined as "usual" received no formal training until week 12, at which time they attended the same training seminar offered initially to the "trained" group. Following each session, objective practical and written assessments confirmed significant workshop "training effects". Analysis of data collected during the administration of 1,012 spirometry tests (by both groups) identified training as the major determinant of an acceptable maneuver. Between week 0 and week 12, a significantly higher frequency of patient tests in the "trained" group satisfied ATS acceptability criteria than in the "usual" group (18.9 % from the "trained" group versus 5.1% from the "usual" group). Once the "usual" group received identical training, results were proportionately consistent with those of the trained practitioners. Although, in general, spirometry performed by neither group met ATS standards, a significant "training effect" was exhibited.

Lung Health Study

Through the incorporation of a comprehensive quality control program, the 1986-1991 Lung Health Study monitored the progress of technician skill level with respect to the production of acceptable spirometric data over a five-year interval (Enright et al, 1991). As a multi-centered, longitudinal clinical trial quantifying within-individual changes in lung function of obstructed smokers, it required accurate and reliable spirometry performance. The subsequent program facilitated a thorough collection of technician expertise/test administration data.

Prior to conducting spirometric tests, all technicians completed 16 hours (distributed over 4 days) of pulmonary function instruction in which spirometry administration comprised one component. Importantly, technician expertise at study commencement was defined entirely by the training acquired during this session (i.e., technicians had no previous spirometry experience). Approximately 9 months into the study a pulmonary function supervisor (pulmonary physician) provided supplementary practical instruction by observing each technician's performance in the field and correcting noted variances in protocol. A realization of continued sub-optimal testing sessions throughout the next nine months resulted in the adoption of a uniform technician quality control program that delivered monthly reports with constructive recommendations to each technician. A test-session-quality rating system based on ATS acceptability and reproducibility criteria, combined, was devised and implemented to quantify the effectiveness of these strategies. It indicated a 44% increase in test quality throughout the duration of the study. For all test sessions, only 2.1 % were deemed unacceptable. Moreover, significant improvements corresponded to the 9 month and 18 month points of intervention inception (Enright et al, 1991). Therefore, a direct correlation between acceptable test session data and technician expertise was recognized.

ii) Study of Test Acceptability Within A Technician Expertise Level

Another multi-centered epidemiologic survey investigating the association between technician expertise and spirometric acceptability originated from the 1991 Swiss Study on Air Pollution and Lung Disease in Adults (SAPALDIA) (Kunzlie et al, 1995). It sought to evaluate the quality (i.e., the acceptability and reproducibility) of spirometry performed in its parent study. Contrary to the Lung Health Study, spirometric quality was not assessed at the time of data collection.

Replicating the spirometric procedures employed in SAPALDIA, technicians at field sites administered spirometry to 15 - 20 new volunteers. However, unlike SAPALDIA, each technician tested all subjects, generating a set of repeated lung function measurements per subject. The variability of FEV_1 and FVC values within these sets was calculated and regarded as a means of assessing technician effects. Observed minimal values indicated consistency across technicians. Furthermore, across field sites, only 10 out of 137 subjects (7.3%) did not fulfil the acceptability criteria in more than one test session, a rate concluded to be unrelated to a specific technician or site.

Investigators attributed the above-mentioned homogeneity to technician expertise. Preceding this study, all technicians completed 3 days of training, 2 months of practice, and one year of field experience in SAPALDIA. Analogous to the Lung Health Study, technicians also received regular supervisory feedback and immediate, computerized acceptability and reproducibility prompts for each spirometric maneuver (Kunzli et al, 1995). Collectively, these initiatives not only supported the technicians' ability to standardize spirometry procedures, but also ensured their equivalency in expertise. An integral component of these standardized procedures was the assessment of spirogram acceptability. Although not directly addressed in this study, consistency in the interpretation of test acceptability was suggested by the reported consistency in test administration. Therefore, such findings propose that technicians of a common skill level interpret spirometric test acceptability similarly.

iii) Study Verifying the Correlation Between Technician Expertise and Test Acceptability

The 1988 United States' National Health and Nutrition Examination Survey of children and adults (NHANES III) demonstrated the influence of rigorous technician training and quality control on the production of acceptable spirometric results. Following one week of formal training to complete a National Institute for Occupational Safety and Health (NIOSH)-approved spirometry course, technicians participated in four pilot studies during which they practiced and received additional supervised monitoring and instruction (Hankinson and Moon Bang, 1991). Once the main study was initiated, all spirograms were collected and reviewed by senior quality control technicians who provided appropriate follow-up advice. Periodic site evaluations supplemented the routine visual inspection of graphic records. At the completion of each test, computerized feedback displayed the maneuver's corresponding curve and indicated its compliance with ATS acceptability criteria. Of the study sample, 4.6% generated fewer than two acceptable curves, a result comparable to other multi-centered studies (Kellie et al, 1987). Therefore, a positive relationship between technician expertise and the interpretation of test acceptability was demonstrated (Hankinson and Moon Bang, 1991).

iv) Compliance with ATS-Established Technician Qualifications and Test Acceptability

In 1984, two occupational epidemiologic studies comparing the respiratory health of machinists and textile workers reported that 15.6% of their study participants generated spirometric results considered to be of poor quality, in part, as a result of failure to meet ATS acceptability criteria (ATS, 1987). Both of these studies conformed to the1982 ATS guidelines for technician expertise and spirometric procedures (Eisen, 1987). These guidelines, though flexible, recommended that technicians possess a high school education and a "strong background in mathematics or biological sciences". They also advised six months of laboratory training under the assumption that instruction for all types of pulmonary function tests would be concurrent (ATS, 1987). In these studies, technician preparation beyond compliance with ATS criteria was not clarified. Also, test acceptability was limited to technician interpretation without supervisory assistance. Although neither study formally investigated the effect of technician expertise on the proportion of acceptable lung function tests, each alluded to their association by acknowledging technician skill as a concern (Christiani, et al, 1985).

v) Technician Expertise in Studies Reporting Poor Quality Spirometric Data

One 1993 occupational epidemiologic study, a respiratory health assessment of sewer and water-treatment workers, identified "low quality spirometric data" among its results. Spirograms from 87 of its 217 subjects (40.3%) were rejected (Richardson, 1995). With the exception of the type of automated spirometry system employed, no reference was made to the spirometric protocol followed; nor were technician qualifications, training, or experience addressed. Consequently, it is unclear why the quality of the spirograms was not questioned during spirometry administration. However, the study's findings underscore the importance of careful test acceptability interpretation at this stage.

Review of the studies above establishes a clear consensus: "The single most significant factor in the production and collection of high quality spirometric data is the person conducting the test" (Clausen, 1982). Thus, because the technician both administers the test and interprets its quality, he/she, through subjective, visual analysis, both ensures correct performance of the maneuver and judges its acceptability (Quanjer, 1993).

1.5.2 Stage 2: Test Interpretation

Once spirometric data are assembled, their acceptability is frequently re-evaluated by a second interpreter (Chusid, 1983). Whether this interpreter is a physician reading the curves for clinical diagnosis, a researcher selecting the "best test"[1] values for database construction, or a senior pulmonary function technician reviewing results for verification of test quality, none are present at the time of test administration. Retrospective epidemiologic studies employing pre-existing or archived records, in particular, exemplify this situation (Tockman and Comstock, 1989). Because spirometric maneuvers cannot be directly observed, analyses, instead, are based on visual inspection (qualitative and quantitative) of submitted spirograms (Quadrelli et al, 1996). Therefore, the interpreter's ability to reliably assess their acceptability becomes a consideration.

At this stage of judging spirogram quality, expertise levels vary from basic familiarity with ATS acceptability criteria to a comprehensive knowledge of respiratory physiology (Ruppel, 1997). Despite a thorough manual and computer search of past literature, only one investigation specifically examining the relationship between training and spirogram acceptability assessment beyond the test administration stage was located. As part of the

[1] Best test curve refers to the curve which produces the largest sum of FEV_1 plus FVC (ATS, 1995)

New Zealand study described in Section 1.5.1, trained and untrained participants were asked to interpret the acceptability of randomly selected spirograms (Eaton et al, 1999). Three additional studies relating expertise to the interpretation of graphic tracings also provided relevant information. Two of these investigations compared pulmonary expertise with the ability to interpret spirometric patterns indicative of pulmonary abnormalities (i.e., obstructive or restrictive defects) (Hnatiuk et al, 1996; Quadrelli et al, 1995). The third study correlated medical expertise with electrocardiogram interpretations. All were inter-rater reliability studies measuring the degree of agreement/disagreement between interpreters who subjectively evaluated graphic records (Westdorp et al, 1992).

i) Study of Variance in the Interpretation of Graphic Records Between Medical Expertise Levels

In the latter phase of the New Zealand study, participating practitioners were asked to interpret 25 randomly selected spirometric records from their individual practices (Eaton et al, 1999). Two experienced pulmonologists (serving as "gold standards") subsequently reviewed 559 previously interpreted spirograms using the same information that was available to the primary care physicians. In only 296 cases (53%) were the practitioners' interpretations evaluated as correct (with no significant difference between the trained and usual groups). From these results it was inferred that accurate interpretation of spirometric tracings required a higher level of pulmonary expertise than could be acquired during basic spirometry training sessions.

ii) Study of Variance in the Interpretation of Graphic Records Within An Expertise Level

The degree of discordance in the assessment of spirograms was the focus of a study conducted at the University of Buenos Aires in 1996 (Quadrelli et al, 1996). Fifteen pulmonologists each determined the presence of respiratory defects in a common set of 15 spirograms. The results indicated 76% maximal agreement (defined as the maximum amount of concordant observations in relation to the total observations) among interpreters. Within the spirogram set, two "problem" curves were included: one with variations higher than 40% between the two best tests (thereby failing to meet ATS reproducibility criteria) and one with initial hesitations and "multiple gaps in the curve development" (thereby failing to meet ATS acceptability criteria). The inclusion of such curves allowed researchers to establish whether pulmonologists' evaluated spirogram quality (i.e., through the application of ATS acceptability and reproducibility criteria) before interpreting the curves for pulmonary abnormalities. To facilitate this assessment, researchers provided all pulmonologists with the option of assigning curves to a "not assessable" category since compliance with ATS criteria requires the exclusion of curves rendered unacceptable from subsequent analyses. For the "problem" spirograms, only 14% of tests with higher than 40% variation among the curves and 33% of those displaying a "grossly imperfect curve" were considered inadequate for interpretation by the pulmonologists. Based on these results researchers concluded that not only "is there substantial disagreement in the interpretation of spirometry" but also "there is a lack of either concern or awareness regarding spirogram quality". Thus, this study indicates that interpretation of graphic records varies within a single level of expertise.

iii) Study of Variance in the Interpretation of Graphic Records Between Expertise Levels

As part of the 1996 Walter Reed Army Medical Centre's Continuous Quality Improvement Project in Washington, D.C., researchers sought to compare interpretations submitted by practicing general internists with those of board-certified pulmonologists (Hnatiuk et al, 1996). Given a series of previously validated (i.e., acceptable and reproducible) spirograms, each was asked to independently identify the presence of restrictive or obstructive patterns. While the pulmonologists referred to the ATS guidelines established for interpreting such patterns, the internists relied on past training and experience. The study's results reported 97% concordance (107 out of 110 tests) within the group of pulmonologists. On comparing their interpretations with the internists, this rate declined to 66.4% (73 out of 110 tests). Further, in 30.9% of the cases (34 out of 110 tests) the internists failed to identify abnormalities or simply noted abnormalities that did not exist. However, spirometric results were interpreted as abnormal by both the internists and pulmonologists in 9.9% of the cases (10 out of 110 tests). Therefore, in approximately 33.4% of all spirometric tests reviewed, the interpretations of the general internists, who likely had less experience with lung function testing, differed (statistically) from those of the pulmonologists (Hnatiuk et al, 1996). These findings suggest that expertise level may be an influential factor in the interpretation of graphic records.

iv) Study of Variance in the Interpretation of Graphic Records Between Expertise Levels (In Medical Disciplines Other Than Pulmonology)

Conclusions similar to those above were reported in literature addressing the concordance of emergency physicians and cardiologists when reviewing electrocardiograms (ECGs). This retrospective cohort study conducted in 1992 at the University of Missouri-Kansas City collected ECGs from discharged patients who visited the Emergency Department (Westdorp et al, 1992). All ECGs were interpreted by both a cardiologist and an emergency physician. Of the 143 ECGs read, an overall discordance of 58% between groups was calculated (Westdorp et al, 1992). No discordance rates for within groups (i.e., within medical disciplines) were reported. These results suggest that the interpretation of graphic records may vary by level of expertise, regardless of medical discipline.

Although studies describing the interpreter and his/her role in spirometric assessments are limited, those cited confirm the need for evaluation of graphic records by a second interpreter. Further, since evidence of an association between expertise level and the quality of graphic evaluations exists, the interpreters' training and experience require consideration.

1.6 Factors Influencing Spirometric Interpretation

In addition to technician and interpreter expertise, literature points to a series of other factors that may influence test acceptability. These factors can be categorized as either equipment-related or participant (subject)-related.

1.6.1 Equipment Factors

i) Computerized Software

Current studies continually refer to the use of spirometric software as a means of assuring spirometric test quality (Enright et al, 1991). Such systems are programmed with advanced algorithms which provide multiple checks for patterns of unacceptable tracings and non-reproducible, tabulated data at the time of test administration. However, through a sub-study comparing spirometers employed in SAPALDIA, investigators detected device errors not otherwise recognized (Kunzli et al, 1995). For example, using two different instruments, they were able to reproduce unreliable results without receiving an error message from the computer. Moreover, a 1990 comprehensive survey of 62 spirometers from 37 different sources worldwide recognized that 25% of all computerized systems exhibited "bugs in their software" (Nelson et al, 1990).

ii) The Damping Mechanism of the Spirometer

There is a clear consensus among pulmonary function equipment technicians that specific properties of spirometric models can introduce significant variance between spirometers classified as the same general type (Nelson et al, 1990). In particular, a device's damping attributes directly affect the appearance and, more critically, the accuracy of the graphic output. To clarify, damping is a characteristic that describes the sensitivity with which an instrument reacts to dynamic changes in input signals. If the output signal temporarily "overshoots" the input signal in response to rapid or abruptly changing flow rates (e.g., a cough), the instrument is considered to be under-damped. If the output signal, instead, approaches the input signal curvilinearly, the instrument is said to be over-damped. For example, a flow-volume curve produced by a device with little or no damping properties will display even small fluctuations in flow while these same fluctuations will be smoothed by a device with a damping mechanism. Regardless, either circumstance results in distortion of the recorded curve, rendering it difficult to interpret for acceptability (Wanger, 1992).

1.6.2 Participant Factors

Participant factors associated with spirometric test quality include age, gender, past spirometry experience, smoking status, and history of pulmonary and/or cardiovascular conditions.

i) Age

According to the results of the NHANES III study, it is suggested that the age of the subject performing spirometry may influence his/her ability to generate 2 acceptable tests. When the percentage of participants between the ages of 18 and 55 who performed more than 2 acceptable maneuvers (97.1%) was compared with that of participants over the age of 55 (93.6%), researchers noted a significant reduction, regardless of gender. These values continued to decrease with increasing age after 55. A similar trend, although not as significant, was observed in participants under the age of 18 (i.e., younger participants produced lower acceptability rates). For participants over the age of 65, 45.1% of those with fewer than 2 acceptable maneuvers simply declined to perform more than 2 trials. In subjects under the age of 18, 26.0% required more than 7 maneuvers to produce 3 of acceptable quality. These results led the study's investigators to conclude that, in general, "younger and older subjects have more difficulty understanding and performing spirometry" (Hankinson, 1991).

ii) Gender

Findings from studies investigating the relationship between gender and test acceptability indicate a slight, but, nonetheless, notable difference in acceptability rates between males and females. In the NHANES study, females were found to produce fewer acceptable maneuvers than males of the same age (Hankinson et al, 1991). The Lung Health Study reported similar conclusions (Variability in FEV_1 values was greater for females than for males) (Enright et al, 1991).

iii) Past Spirometry Experience (Learning Effect)

Some studies have reported the presence of a learning or practice effect in spirometry since, on average, results obtained during second visits, usually within weeks (i.e., a time period in which no major biological changes in the study population are likely), are higher than those obtained at first visits (Becklake, 1993). For instance, one study involving healthy non-smokers revealed that variability in FEV_1 values[1] decreased significantly (i.e., 50 ml) with additional testing. After "controlling for all other sources of variation", researchers attributed this observation to the appearance of a learning effect (Burrows, 1986). In addition, spirometric results from the Lung Health Study described a decline in the number of trials required to achieve 3 acceptable maneuvers between consecutive testing sessions. Testing sessions were, on average, 25 days apart. Although the difference proved to be statistically non-significant, researchers still acknowledged it as indicating a learning effect (Wise et al, 1995).

iv) Smoking Status

The relationship between smoking status and acceptable spirometric performance appears ambiguous. While pulmonary function manuals frequently state that smoking affects the subject's ability to take a deep breath and, therefore, perform an acceptable maneuver, results from two community based studies reported that test failure occurred less frequently in smokers than in nonsmokers (Krzyzanowski et al, 1988; Eisen et al, 1987). Increasing evidence of this phenomenon referred to as the"healthy smoker effect", especially in younger populations, suggests that individuals who smoke have higher baseline levels of lung function, and, possibly, less reactive airways than those who do not (Becklake, 1990).

[1] The rationale for using change or variation in FEV_1 values as a means to assess test acceptability is based on the notion that the FEV_1 value is an effort-dependent test, with the largest value representing the greatest effort. Stability in this value over repeated maneuvers indicates consistency in subject effort and, in turn implies test validity (Wise et al, 1995).

v) History of Past Pulmonary and/or Cardiovascular Conditions

Results from various occupational, cross-sectional studies have concluded that test failure (i.e., production of an unacceptable, non-reproducible maneuver) may itself be an indicator of poor respiratory health (obstructive lung disease, in particular) and, thus, exclusion of corresponding subjects potentially creates a selection bias (Becklake, 1990). For example, in an investigation of Pennsylvania railroad workers, individuals with chronic bronchitis produced a greater number of unacceptable maneuvers than did those reporting no respiratory symptoms (Eisen et al, 1985). Also, a study of coal miners related reports of wheezing, shortness of breath, and chronic cough to higher levels of test failure (Kellie et al, 1987). Clinically, recent acute lower respiratory illness is often associated with performance of unacceptable maneuvers (Ruppel, 1997). However, results from the Lung Health Study of smokers diagnosed with mild to moderate pulmonary obstruction indicated a 2.1% test failure rate, a value significantly lower than that observed in other studies (presenting test failure rates between 8% to 20%). Thus, researchers from this study argued that "while we do not dispute these associations, we believe they should not be used as an excuse for poor quality test sessions". Further, they recommended that "better technician training and monitoring be tried before one concedes that test failure is as likely to reflect ill health as it is to reflect incompetence of the technician" (Enright et al, 1991).

1.7 <u>Summary</u>

Literature describes an extensive use of spirometric testing in pulmonary function assessment. To satisfy clinical, occupational, and epidemiologic objectives of lung health diagnosis, treatment, and monitoring, the interpretation of graphic records produced during testing is a primary concern. It must be both reliable and accurate. The American Thoracic Society attempted to assure such validity through compliance with their acceptability and reproducibility criteria. Implementation of these criteria require quantitative and qualitative assessments of the spirometric tracings. Although much of the analysis is computer-generated, studies have indicated the need for subjective judgement in order to confirm acceptability. This judgement is completed twice during the evaluation process, once at the time of test administration while in the presence of the subject and again, retrospectively, in the absence of the subject. Studies also concur that the quality of these interpretations is correlated with the expertise level of the interpreter. Furthermore, because respiratory, epidemiologic studies, in particular, employ multiple interpreters with various credentials, the importance of thoroughly investigating the association between pulmonary expertise and the interpretation of spirometric test acceptability is magnified. The current study attempted to clarify this association.

CHAPTER TWO
MATERIALS AND METHODS

This chapter outlines the study's research objectives and describes the methodology (i.e., data collection procedures and subsequent statistical analysis) utilized to address them.

2.1 Study Objectives and Hypotheses

This study's four primary objectives were:

1. To determine whether inter-rater reliability (raters' agreement/disagreement) for interpretation of spirogram acceptability (based on American Thoracic Society 1994 Criteria) varied according to the raters' levels of respiratory expertise (training and experience).

2. To quantify differences in rater agreement/disagreement for spirograms produced by highly trained technicians and minimally trained technicians within a field setting.

3. To investigate the influence of raters' respiratory expertise on the type of explanations provided for spirograms evaluated as unacceptable.

4. To identify characteristics of test participants (subjects) whose spirograms were recognized as unacceptable.

It was hypothesized that the degree of agreement among raters will depend upon the similarity of their respiratory expertise since deciphering test acceptability is largely a subjective process. Further, it was expected that, because the quality of spirometric tracings is directly related to the performance of the examination, itself, there will be a greater level of both agreement and acceptability for spirograms produced by more highly trained and experienced technicians regardless of the rater's level of expertise.

This study also hypothesized that certain characteristics of the participants performing the spirometric maneuvers may influence the acceptability of spirometric results. These participant factors include:

- Age
- Gender
- Smoking status
- History of pulmonary and cardiovascular conditions
- Past spirometry experience
- Lung function parameters (as percent of predicted values[1])

By examining the inter-rater reliability (degree of agreement) between groups of spirometric interpreters with various levels of expertise, this study attempted to establish the appropriate type of respiratory training and experience required to properly scrutinize spirograms for test acceptability. Moreover, it sought to assess the application and sufficiency of ATS criteria under field conditions.

Realization of the study's objectives, collectively, will optimize the quality and, in turn, maximize the utility of data collected in future pulmonary epidemiologic studies.

2.2 Study Design

This study involved the assembly of two sets of spirograms and their subsequent interpretations by categories of raters with various levels of respiratory expertise (i.e., familiarity with spirometry, pulmonary physiology, and pulmonary epidemiology).

[1] Interpretation of lung function results involves a comparison of observed values with reference or predicted values. These "normal" values are derived from studies of well-defined, healthy populations. Each interval or range of "normal" values represents typically 95% of the sample populations for a specific sex, height, age, weight, and race. The percent of predicted value is determined by dividing the observed value by the reference value and multiplying by 100 (Wanger, 1992).

While both sets of spirograms were compiled to address a common series of basic research questions, each utilized a distinct spirogram collection strategy. Therefore, a comparison of corresponding results for the two sets facilitated an investigation of possible methodological influences on spirogram acceptability and inter-rater reliability.

2.3 Assembly of Spirogram Sets

Each data set was defined according to spirogram origin. Spirograms generated specifically for this study comprised the Primary Data Set[1]. In contrast, the Secondary Data Set[2] partially included archived spirograms derived from previous respiratory epidemiologic research.

2.4 The Primary Data Set

Meeting the study's objectives required interpretation of spirometric tests compiled outside of an established clinical environment under conditions in which population-based epidemiologic studies assessing overall adult respiratory health are often completed. Therefore, spirogram collection procedures for this data set, including participant sampling and equipment considerations, were designed to reflect a typical field setting.

2.4.1 Selection of Study Group Performing Spirometric Tests

i) *Study Population*

Pulmonary function surveys conducted for screening purposes frequently examine a cross-sectional sample of the general population thereby exposing technicians directing lung function tests to a broad spectrum of subjects. To ensure inclusion of spirometric results exhibiting similar variability, the participant recruitment approach employed in

[1] As used here, 'Primary data' refers to new data collected specifically to address the present study's research questions.

[2] As used here, 'Secondary data' refers to pre-existing data initially compiled to address a different set of research questions.

this study targeted the general population.

ii) *Recruitment of Study Group*

In collaboration with the Alberta Lung Association and the Alberta Asthma Centre, study personnel organized, advertised (Appendix 1) and operated a lung health awareness clinic. This interactive clinic presented comprehensive information regarding the recognition, prevention, and management of respiratory conditions. A display which introduced spirometry as a fundamental, pulmonary function assessment tool and promoted the availability of complimentary testing was featured (Appendix 2).

a) Location of Clinic

When considering possible clinic venues, ease of access to a study group representative of the general population was a primary concern. Because shopping malls provide convenient, essential services and, therefore, attract a consistently large volume of people, they were recognized as suitable locations. Moreover, according to information obtained from a local market research consulting firm, the specific mall selected for the study caters to a notably diverse clientele which best parallels the general population (personal correspondence, Thompson-Dobo).

b) Timing of Data Collection

Lung function testing was offered over a 5-day period during the following times:

Tuesday through Friday: 4:00 p.m. to 9:00 p.m.

Saturday: 10:00 a.m. to 5:00 p.m.

These times satisfied two study requirements:

1. Compliance with a pre-calculated period of time sufficient for gathering study data.
2. Incorporation of testing times which afforded the general population an opportunity to participate, minimizing selection bias.

iii) *Study Group Sampling*

Visitors expressing an interest in completing spirometry were first approached by a research assistant who determined their eligibility to participate in the study and explained that lung function testing would be offered and, subsequently, administered regardless of study involvement. The research assistant then met with each potential volunteer individually to distribute information letters (Appendix 3), outline and clarify study details, and address any emerging questions or concerns. Lastly, consent forms (Appendix 4) were signed and completed by those who agreed to enter the study.

iv) *Inclusion and Exclusion Criteria*

Each visitor's participation status was evaluated using the following criteria.

a) Inclusion Criteria:

1. The participant was 18 years of age or older.
2. The participant was able to understand verbal English instructions.
3. The participant could provide written, witnessed informed consent.

b) Exclusion Criteria:

The participant presented contraindications to the spirometric test, itself.

Such contraindications included:

1. Acute illness that could interfere with test performance (e.g., nausea and vomiting)
2. Recent myocardial infarction or pulmonary emboli*
3. Recent abdominal or thoracic surgery*
4. Recent eye surgery*

All visitors meeting inclusion criteria were invited to participate.

Once again, access to lung function testing was not dependent upon study involvement.

*For liability reasons, spirometric testing could not be offered to any visitor with these restrictions.

2.4.2 Data Gathering Procedures

Data collection consisted of two phases:

Phase 1. Questionnaire administration

Phase 2. Pulmonary function testing

Phase 3. Sampling of Spirograms

i) Phase 1. Questionnaire Administration

Each participant completed a brief, interviewer-administered questionnaire which included such items as:

- Date of birth
- Gender
- Previous experience with spirometry
- Smoking history (pack-years)

While most questions originated from a well-validated, standardized, respiratory health questionnaire[1] (IUATLD, 1986), those relating to previous spirometry experience were constructed specifically for this study. However, all questions followed a similar format preserving structural simplicity and cohesiveness (Appendix 5).

Research assistants received explicit verbal instructions regarding questionnaire administration. Further, periodic monitoring of both completed questionnaires and the interview process itself, ensured systematic collection of information and adherence to the study protocol.

Before proceeding to spirometry, each participant was assigned a confidential identification number. These identification numbers linked questionnaire responses with corresponding lung function results. They also distinguished the sequence in which participants completed lung function testing. This information was important for assessing technician performance throughout the study period.

[1] Questions were adapted from the European Respiratory Health Survey Questionnaire (IUATLD, 1986).

ii) Phase 2. Pulmonary Function Testing

a)Technician Training and Experience

To examine the relationship between technician training and the production of acceptable spirograms, all pulmonary function tests were administered by two technicians whose training and experience differed (Table 2.1).

These two levels of technician qualifications reflect the range of respiratory expertise found in pulmonary epidemiologic research. In particular, the minimally trained technician received spirometry instruction modeled after a previous field study's protocol. Alternately, credentials outlined for the highly trained technician reflected those required by studies accessing clinical facilities.

Table 2.1 Summary of Technician Training and Experience

Respiratory Expertise[1]	**Technician 1**	**Technician 2**
Training	Qualifications • Received a total of 10 hours of individualized, formal spirometry instruction (Appendix 6) • High school education • 1 year of university in physical, biological, and mathematical sciences	Qualifications • Registered Certified Pulmonary Function Technician (documenting competency in lung function testing) • High school education
Experience	No previous spirometry experience	Minimum of 15 years of spirometry experience
Classification	Minimally Trained Technician	Highly Trained Technician

[1]. The ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories recommends, but does not mandate, that technical staff conducting pulmonary function tests receive 6 months of supervised training. This 6 month period was developed with the consideration that training for all tests performed in the pulmonary function laboratory, not just spirometry, would be concurrent. In addition, the committee suggests completion of high school education and 1 year of college-level courses in the biological and physical sciences by all technical staff. For supervisory staff, pulmonary function credentials granted by relevant professional bodies, as well as 2 years laboratory training and experience, is advised (Gardner, 1983).

b) Equipment Considerations

This study employed two portable, pneumotach flow-sensing SpiroSense© spirometers equipped with computerized software satisfying ATS recommendations. Prior to their use, the validity (i.e., precision and accuracy) of both devices was confirmed by instrument technicians in an established pulmonary function laboratory. The spirometers were calibrated each day before testing and after sensor replacement using a standardized 3 litre calibration syringe. Routine maintenance at the conclusion of each testing period included disassembly and proper cleaning of mechanical components.

Although identical in make, model, and year, the spirometric systems were rotated daily between technicians to eliminate possible equipment bias.

c) Pulmonary Function Testing Procedures

All participants completed two consecutive spirometry sessions, each directed by a different technician. To ensure participant privacy and prevent technicians from interacting with each other, sessions were conducted in separate cubicles. The technician seen first by a participant was based on availability at the time. However, logs of technician order were kept to permit investigation of any variance in the anticipated participant learning effect that could be a consequence of this order.

1) Participant Preparation

Prior to directing the spirometric maneuver, each technician recorded the participant's age, gender, race, height and weight[1], all physical characteristics applied to prediction equations which provide the context for evaluating pulmonary function results. A brief pulmonary and cardiovascular health history (Appendix 7) was also compiled using a questionnaire[2] automatically incorporated within the spirometric software package[3].

The reliability of all responses was evaluated using a test-re-test approach whereby sets

[1] In compliance with standardized procedures, participants removed their shoes before height and weight were measured.

[2] Questions were adapted from the American Thoracic Society's 1978 Adult Questionnaire (ATS, 1979).

[3] Spirometers were operated with SpiroSense V2.09 software.

of answers to questions repeated at different times were compared for similarity. As previously mentioned, all participants were tested by both technicians, individually. Therefore, they proceeded through the participant preparation phase twice, answering duplicate questionnaires. Subsequent comparison of responses to identical questions determined their consistency and, in turn, inferred their accuracy[1]. Lastly, those participants whose paired answers revealed discrepancies were contacted by telephone for clarification of correct information.

2) Maneuver Performance

After the technician explained and demonstrated the spirometric maneuver, each participant was coached through repeat efforts (to a maximum of 8 attempts) until 3 trials were acceptable. Importantly, technicians[2] judged the acceptability of each trial despite availability of interpretation algorithms within the computerized spirometric system which provided automatic, immediate prompts appraising test performance.

At the conclusion of a test session, technicians saved and stored maneuvers they judged as acceptable.

3) Interpretation of Participants' Results

Only the certified pulmonary function technician reviewed results with participants. Interpretation was limited to a discussion of whether or not values were within the "normal range" in relation to the applied set of reference values. However, a pulmonary physiologist examined all of the spirograms at the conclusion of the testing period and identified participants who should be contacted and referred to an appropriate health care professional[3].

[1] While reliability does not address the issue of accuracy, responses which are not reliable are unlikely to be accurate.

[2] Technicians were blinded to each other's results for the same participant. To clarify, the technician who conducted the participant's second spirometry session was unaware of the results obtained by the first technician.

[3] One (1.0%) participant produced abnormal results and, therefore, required contact.

iii) Phase Three. Sampling of Spirograms

Two spirograms from those the technician stored for each participant were selected to form the Primary Data Set. These spirograms represented maneuvers that generated the two largest sums of FEV_1 plus FVC, a calculation used conventionally in "best test" selection (ATS, 1995). Further, provision of two spirograms allowed raters the option of comparing curve shape pattern, a factor frequently considered when ascertaining the origin (e.g., physiological or technical) of a perceived, questionable anomaly. As a result of test administration by both technicians, each participant contributed four spirograms to the Primary Data Set, two to each subset classified by technician expertise (Figure 2.1).

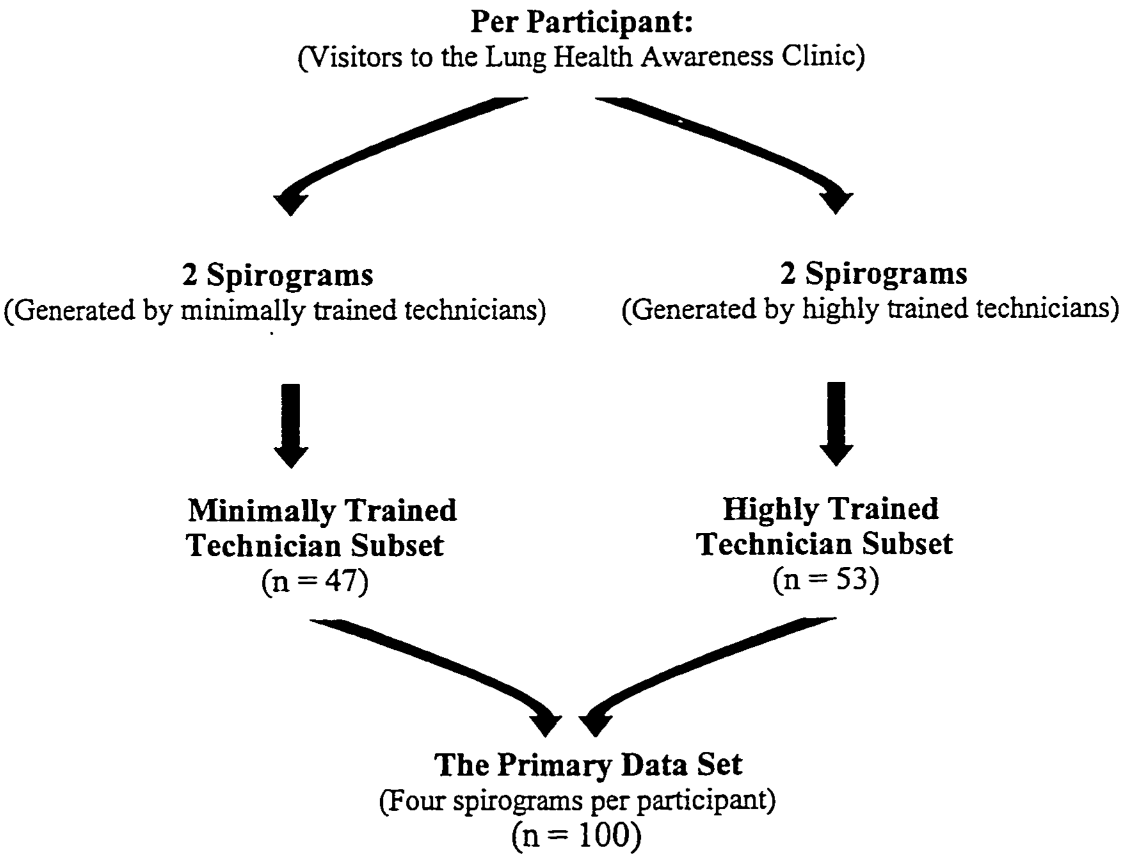

Per Participant:
(Visitors to the Lung Health Awareness Clinic)
2 Spirograms
(Generated by minimally trained technicians)
2 Spirograms
(Generated by highly trained technicians)
Minimally Trained
Technician Subset
(n = 47)
Highly Trained
Technician Subset
(n = 53)
The Primary Data Set
(Four spirograms per participant)
(n = 100)

Figure 2.1 Flowchart of spirogram selection process for the Primary Data Set

Note: n refers to the number of participants

2.5 The Secondary Data Set

Analogous to the Primary Data Set, this data set consisted of spirograms categorized into two subsets based upon technician expertise. However, the inclusion of archived or pre-existing spirograms required implementation of a distinct methodology for each subset's construction.

2.5.1 Construction of the Minimally Trained Technician Subset

This subset consisted of archived spirometric results, exclusively. Spirograms were collected from a previous epidemiologic respiratory health assessment survey[1] conducted in a remote community where certified pulmonary function technicians were not available for the study.

i) Description of Respiratory Health Assessment Survey

a) Study Population

All adult community members who volunteered to participate formed the study population. These participants responded to media advertisements explaining the study and requesting volunteers.

b) Pulmonary Function Testing

1) Technician Training and Experience

All personnel completed basic spirometric training prior to data collection. The type and amount they received served as a template for the instruction[2] delivered to the minimally trained technician in the Primary Data Set. Therefore, the respiratory expertise of the technicians for both data sets was considered equivalent.

2) Equipment

Technicians used two pneumotach, flow-sensing MultiSpiro© spirometers connected to

[1] Spirograms were collected from the Adult Lung Health Component of the Fort McMurray Alberta Oil Sands Community Exposure and Health Effects Assessment Study.

[2] Minimally trained technicians for both data sets received spirometric training by the same instructor.

portable computers[1]. Documentation confirming the validity of this equipment (i.e., maintenance inspection prior to study commencement) was not available. Nevertheless, devices were reportedly calibrated at the beginning of each testing period.

3) Pulmonary Function Testing Procedures

Five consecutive days of lung function examinations were completed according to the protocol established by the Respiratory Health Assessment Survey. At the end of each spirometry session, technicians saved and stored all acceptable maneuvers.

ii) Sampling of Archived Spirograms

a) Spirogram Selection Criteria

Archived spirograms were selected to satisfy three criteria.

1. Spirograms chosen were generated from each participant's first spirometry session. This restriction was implemented in an attempt to limit the potential bias created by each participant's prior spirometry experience, thus reducing the possibility of a learning effect confounding the association between technician expertise and spirogram acceptability.

2. Spirogram selection was dependent upon technician experience, a factor hypothesized as affecting the quality of spirometric results. In the Respiratory Health Assessment Survey, not all technicians conducted an equal number of spirometry tests. Of twelve technicians involved, seven administered the majority of the tests to an approximately equivalent number of participants. To standardize the level of technician experience, spirograms were collected from sessions directed by one of these seven technicians.

3. Using criteria applied to the Primary Data Set, two spirograms producing the

[1] Computers were equipped with MultiSpiro Spirometric Software.

two largest sums of FEV_1 plus FVC were selected from within each of the previously identified spirometric sessions.

b) Participant Information

Participants whose spirograms contributed to the subset became the study population. Accessible, descriptive information regarding participants was limited to their spirometric results and their responses to a concise questionnaire regarding smoking status. Therefore, in addition to age, gender, race, height, and weight (all factors indicated for calculation of percent predicted lung function values), records included only a brief smoking history.

2.5.2 Construction of the Highly Trained Technician Subset

Assembly of this subset required both the production and collection of spirograms generated by highly trained and experienced technicians under conditions consistent with those encountered by personnel in the Respiratory Health Assessment Survey.

i) Selection of Study Group Performing Spirometric Tests

a) Study Population

A study group similar to that involved in the Respiratory Health Assessment Survey was required. Past research identifying factors that influence the production of acceptable spirograms has recognized participant age as a significant determinant (Hankinson et al, 1991). Thus, inclusion of a study group featuring a comparable age distribution became the fundamental consideration.

The Respiratory Health Assessment Survey comprised participants primarily between the ages of 20 and 55. Because faculty, staff, and students in the investigator's department at the University of Alberta represented a similar age range, they were regarded as a suitable source from which to recruit this subset.

b) Recruitment of Study Group

Via department mailboxes and electronic mail, faculty, staff, and students received letters introducing and describing the study (Appendix 8). They were then contacted by a follow-up phone call within the next week and asked if they would like to participate.

c) Study Group Sampling

The eligibility status of potential participants (i.e., departmental members who, upon contact, agreed to become involved) was assessed using the set of inclusion and exclusion criteria previously applied to participants who comprised the Primary Data Set's study group (Section 2.4.1).

ii) Data Gathering Procedures

To replicate the methodology outlined in the Respiratory Health Assessment Survey, participants answered a set of identically phrased, technician-administered questions (Appendix 9) regarding smoking status prior to performing spirometry.

a) Questionnaires

Items incorporated within the questionnaire were limited to:

- Date of birth
- Gender
- Smoking history

With the exception of these items, no information pertaining to participants in the Respiratory Health Survey was accessible. Thus, in contrast to the Primary Data Set's assembly, questions regarding past spirometry experience and history of pulmonary and/or cardiovascular conditions were not asked.

Once again, the technician assigned each participant a confidential identification number to facilitate linkage of questionnaire data with corresponding lung function data.

b) Pulmonary Function Testing

1) Technician Training and Experience

A certified, pulmonary function technician with an expertise level resembling that of the technician employed in the Primary Data Set was responsible for test administration.

2) Equipment

To account for any equipment-related differences that might affect interpretation of spirograms (e.g., nature of the print-out), all tests were performed utilizing spirometers from the Respiratory Health Assessment Survey.

3) Pulmonary Function Testing Procedures

The technician entered each participant's age, gender, race, height and weight[1] directly into the computer's database for subsequent calculation of percent of predicted lung function values. Once they received detailed instructions from the technician, participants were coached through spirometric maneuvers. At the end of each test session, all maneuvers judged acceptable by the technician were saved and stored.

The technician described the spirometric results, in general terms, to each participant. A pulmonary physiologist was consulted regarding findings considered to be of questionable clinical significance.

c) Sampling of Spirograms

Using a selection strategy consistent for all data sets, a pair of spirograms representing a participant's two "best test" curves was extracted from those stored for each test session. (Figure 2.2).

[1] Heights and weights were measured by technicians after participants removed their shoes.

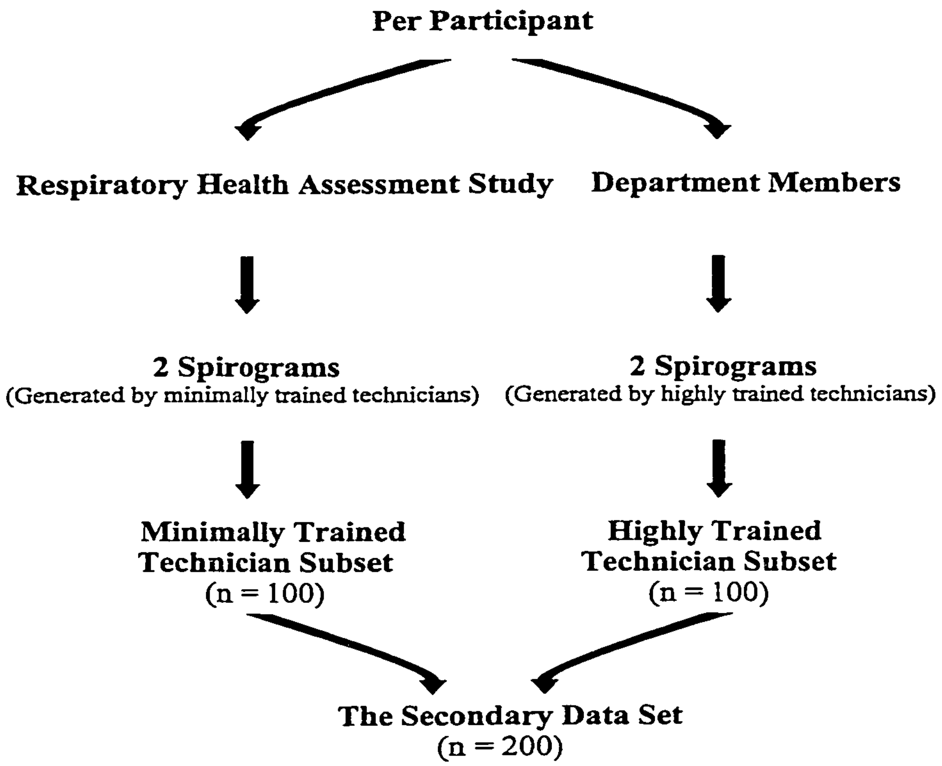

Per Participant
Respiratory Health Assessment Study
Department Members
2 Spirograms
(Generated by minimally trained technicians)
2 Spirograms
(Generated by highly trained technicians)
Minimally Trained Technician Subset
(n = 100)
Highly Trained Technician Subset
(n = 100)
The Secondary Data Set
(n = 200)

Figure 2.2 Flowchart of spirogram selection process for the Secondary Data Set

Note: n refers to the number of participants

d) Equalization of the "Practice Effect" Between Subsets

Once again, participants who generated spirograms for the Highly Trained Technician Subset each completed one spirometric test session. Since the quality of spirometric output is dependent upon the participant's proficiency or skill (i.e., a learning effect), only spirograms derived from the participants' first test sessions in the Respiratory Health Assessment Survey were considered, thus equalizing the effect of practice between subsets (Section 2.5.1 ii a).

2.6 Variations Between the Two Data Sets

The two, principal, methodological differences between data sets were the type of spirometric apparatus employed and the quantity of information available for each participant (Table 2.2).

Table 2.2 Summary Comparison of Methodological Components Between Data Sets

	Primary Data Set		Secondary Data Set	
Methodological Component	Minimally Trained Technician Subset	Highly Trained Technician Subset	Minimally Trained Technician Subset	Highly Trained Technician Subset
Study group	Visitors who received spirometric testing at the Lung Health Awareness Clinic		Participants involved in a Respiratory Health Assessment Survey	Faculty, staff, and students from a university department
*Participant information from questionnaires	Age, gender, smoking status, past spirometry experience, history of pulmonary and/or cardiovascular conditions		Age, gender, and smoking status	
Lung function testing:				
-*Equipment	2 Portable, pneumotach, flow-sensing SpiroSense© spirometers (damping characteristics)		2 Portable, pneumotach, flow-sensing MultiSpiro© spirometers (without damping characteristics)	
-ϕTechnician expertise	Minimally trained and experienced technician	Highly trained and experienced technician	Minimally trained and experienced technician	Highly trained and experienced technician
- Sessions completed per participant	1 session	1 session	5 sessions (1 baseline and 4 test days)	1 session
Selection of two contributing Spirograms	Pair generating the 2 largest sums of FEV_1+ FVC	Pair generating the 2 largest sums of FEV_1+ FVC	From session 1: Pair generating the 2 largest sums of FEV1+FVC	Pair generating the 2 largest sums of FEV_1+FVC

* Denotes methodological components which differed between data sets

ϕ Based on definitions of technician expertise in Table 2.2

2.6.1 Spirometric Apparatus

Regardless of data set, all pulmonary function testing utilized portable, flow-sensing, pneumotach spirometers. However, the specific make of each device and, in turn, it's sensitivity (i.e., damping characteristics), differed between sets.

2.6.2 Participant Information

Archived spirograms for participants contributing to the Secondary Data Set provided no information concerning presence of pulmonary and/or cardiovascular conditions or previous experience with spirometric testing; yet, both were acknowledged as important potential factors affecting the quality of spirometric results. In order to study their influence, participants included in the Primary Data Set answered questions assessing both factors. Subsequent statistical analysis of these factors was simply restricted to the Primary Data Set.

2.7 Rationale for Inclusion of Two Data Sets

The inclusion of both data sets enabled an examination of two important issues. First, since the nature of the recorded graphic output is, in part, a reflection of the spirometer's sensitivity (i.e., damping characteristics), it was hypothesized that the spirometer, itself, may affect the interpretation of acceptability. Therefore, employment of two different models facilitated an investigation of the equipment influence on interpreter agreement/disagreement for spirograms between data sets. Secondly, comparison of results between data sets served as a means by which statistical findings could be validated and assessed for generalizability.

2.8 Number of Spirogram Pairs in Each Data Set (Sample Size Considerations)

One of the primary objectives of this study was to determine whether inter-rater reliability for interpretation of spirogram acceptability differed according to the raters' levels of respiratory expertise. Consistent with most inter-rater reliability studies, analysis utilized the kappa statistic which measures the degree of agreement among raters in excess of that expected by chance, alone (Fleiss, 1981). Therefore, the following sample size equation for kappa was applied in order to estimate the number of spirogram pairs to include in each subset (Norman, 1994):

$$N = \frac{z^2_\alpha p_o(1-p_o)}{\delta^2(1-p_e)^2}$$

where: z_α = value of the standard normal distribution corresponding to a significance level of alpha (z is 1.96 for a two-sided test using a Type I error (alpha level) of 0.05)

p_o = proportion of observed agreement

p_e = proportion of expected agreement by chance alone

δ =confidence interval around estimated kappa (for distinguishing kappas that differ by 0.2, the required confidence interval would be ± 0.1 around the estimates)

Rationale for selecting a confidence interval of 0.2:
To maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics, the following labels are conventionally assigned to each corresponding range of kappa (Landis and Koch, 1977):

Table 2.3 Categorization of Kappa Values

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Note that each category is based upon a kappa interval of 0.2. Thus, a confidence interval of 0.2 (i.e., ± 0.1) was considered to be an appropriate estimate for the sample size calculation.

A series of two-by-two tables was then constructed to calculate several possible p_e values under the assumption that all levels of interpreters involved in the comparison independently accepted and, therefore, rejected the same number of spirogram pairs. Possible p_e values were calculated using the method outlined below:

Let the data be expressed in the following manner (Table 2.4):

Table 2.4 Standard Two by Two Table Design

		Interpreter 1		Marginal
		Acceptable	Not Acceptable	proportion
Interpreter 2	Acceptable	a	b	p1
	Not acceptable	c	d	q1
	Marginal Proportion	p2	q2	1.0

a = proportion of spirogram pairs classified as acceptable by both interpreters
b = proportion of spirogram pairs classified as acceptable by interpreter 2 and as not acceptable by interpreter 1
c = proportion of spirogram pairs classified as acceptable by interpreter 1and as not acceptable by interpreter 2
d = proportion of spirogram pairs classified as not acceptable by both interpreters

None of the actual values for a,b,c, or d were known prior to the start of the study. The formula below was used to calculate the expected proportion of interpreter agreement due to chance (p_e):

$$p_e = a + d$$

Thus, calculation of p_e required values for a and d. A series of theoretical marginal proportions was used to derive a set of numbers corresponding to a and d.

The following pairs of values for the marginal proportions, p1, p2, q1, and q2 were substituted into the above two by two table:

p1 and p2	q1 and q2
0.20	0.80
0.30	0.70
0.40	0.60
0.50	0.50

From these tables, a and d values were obtained, permitting calculation of p_e values. Once p_e values were established, a pair of p_o values, one a difference of 0.1 and the other a difference of 0.2 from each p_e value, was applied to the above sample size formula. Values for N derived from each scenario were summarized and examined (Table 2.5).

Table 2.5 Summary of Sample Size Calculations Corresponding to Different Pairs of p_e and p_o values

p_e (Expected proportion of agreement)	p_o (Observed proportion of agreement)	N (# of spirogram pairs required)
0.68	0.78	161
	0.88	99
0.58	0.68	118
	0.78	93
0.52	0.62	98
	0.72	84
0.50	0.60	92
	0.70	81
Mean value for N		103

The required number of spirogram pairs in each subset was estimated to be approximately one hundred. With the implementation of the sampling strategies discussed for each subset's study group, an appropriate sample size was achieved. In particular, for the Primary Data Set, the first one hundred participants who completed two pulmonary function test sessions (i.e.,1 session per technician) contributed their results to the simultaneous assembly of both subsets. With respect to the Secondary Data Set, pairs of archived spirograms from the first one hundred participants who completed spirometric tests in the Respiratory Health Assessment Survey comprised the Minimally Trained Technician Subset. Similarly, the Highly Trained Technician Subset was formed using spirogram pairs from the first one hundred departmental members who received pulmonary function testing.

2.9 Preparation of Spirograms for Interpretation

Through the application of a standardized structural format, separate spirogram interpretation forms corresponding to each spirometric tracing were created (Appendix 10). With the exception of the three components necessary to judge acceptability [i.e., the flow-volume loop, volume-time curve, and lung function values (including percent of predicted)], all information presented on the original spirograms was removed. Identification numbers which replaced the participants' identities were assigned to interpretation forms derived from spirogram pairs. For each data set, pairs of spirogram interpretation forms comprising subsets were merged together and randomized using a random numbers table. Assignment of a second group of identification numbers for the combined, randomized set blinded raters to technician expertise level.

In summary, two randomized sets, each containing 400 paired spirogram interpretation forms, were devised.

2.10 Interpretation of Spirograms

To examine the relationship between respiratory expertise and interpretation of spirogram acceptability, inclusion of raters representing a suitably broad spectrum of respiratory backgrounds was required.

2.10.1 Categorization of Raters

Four groups of interpreters were defined and constructed based upon their level of pulmonary expertise. Qualifications of raters in each category either correlated with those recommended for the various strata of staff employed in pulmonary function laboratories or with those of personnel involved with respiratory epidemiologic research (Table 2.6).

2.10.2 Number of Raters Comprising Each Group

The targeted number of raters per respiratory expertise category was three. In the case of a possible discrepancy between two raters, inclusion of a third rater served to clarify the interpretation which best reflected their respective expertise level. However, due to the study's time constraints and the limited number of accessible pulmonary specialists and respiratory epidemiologists, it was acknowledged that ascertainment of three raters may not be feasible. Therefore, two raters per group was deemed acceptable.

Table 2.6. Comparison of Described Categories of Raters with ATS-Recommended Qualifications

Rater Group	Description of Raters	ǂATS-Recommended Qualifications
Pulmonary Specialists (n = 2)	Pulmonary Physiologists (PhD)or Pulmonary Physicians (FRCPC)	ATS- recommended qualifications of the Medical Director in the Pulmonary Function Laboratory
Respiratory Epidemiologists (n = 2)	Epidemiologists (PhD) specializing in respiratory health studies	No corresponding ATS category of personnel qualifications
Certified Respiratory Technicians (n = 4)	Certified Pulmonary Function Technicians or Respiratory Therapists with at least 2 years of post-graduate practical experience	ATS-recommended qualifications of Supervisory Staff in a Pulmonary Function Laboratory
Non-certified, Minimally Trained, Respiratory Research Assistants (n = 3)	*Research Assistants with no formal respiratory training but with knowledge of ATS Standardization of Spirometry Document (1994 Update)	No corresponding ATS category of personnel qualifications Less than recommended for technical staff

* Holding Bachelor of Science Degrees in Biological and Mathematical Sciences
ǂ Categories of Personnel

(Gardner, 1983)

2.10.3 Selection of Raters

Selection of raters varied with interpreter category.

i) Pulmonary Specialist Group

Pulmonary physician and/or physiologist selection was based upon the directory of university faculty members in the Division of Pulmonary Medicine who were associated with either the hospital's Pulmonary Function Laboratory, itself, or with other pulmonary research projects utilizing spirometry as part of their methodology.

ii) Respiratory Epidemiologist Group

The university's only respiratory epidemiologist was asked to both participate and provide the name of a colleague with similar expertise.

iii) Certified Respiratory Technician Group

Four pulmonary function technicians and respiratory therapists administering spirometric tests at the University Hospital's Pulmonary Function Laboratory were approached. It is important to note that, although staff from this laboratory were also involved in lung function testing for construction of both data sets, no technician interpreted tests he/she administered. Specifically, two of these technicians read only half of the spirograms.

iv) Non-certified, Minimally Trained, Respiratory Research Assistant Group

Three departmental research assistants, experienced in data compilation and management of public health research studies, were selected. Additionally, each assistant was familiar with the ATS guidelines (i.e., read and reviewed the ATS Standardization of Spirometry 1994 Update) for evaluating spirogram acceptability (ATS, 1994).

2.10.4 Recruitment of Interpreters

Regardless of interpreter group, all potential raters were sent information letters outlining the study (Appendix 11). They were then contacted via a follow-up telephone call requesting their participation.

For reasons previously stated, (Section 2.10.2) the Pulmonary Specialist Category and Respiratory Epidemiologist Category each contained only two raters.

2.10.5 Data Gathering Procedures

i) Spirogram Interpretation Form Packages

Raters agreeing to participate each received a package containing the following items:

- A personalized letter providing instructions for completing interpretation forms (Appendix 12)

- A consent form to be signed and returned at the rater's earliest convenience (Appendix 13)
- Two sets of 400 interpretation forms (Appendix 10)
- A copy of the American Thoracic Society's Standardization of Spirometry 1994 Update for reference (Appendix 14)

ii) Spirogram Interpretation Procedures

Each interpreter was asked to independently judge only the acceptability (i.e., not reproducibility) of tracings displayed on spirogram interpretation forms and, if applicable, answer one of two supplemental, open-ended questions (Appendix 10). Further, raters were blinded to the findings of all other raters.

Upon return of completed packages, identification numbers were assigned to individual raters, ensuring their confidentiality.

2.11 Data Management

2.11.1 Coding of Responses to Open-ended Questions on Interpretation Forms

Development of an appropriate coding scheme required four steps. Initial response categories were first created following a thorough review of raters' comments. Each comment, transcribed verbatim, was then grouped into one of these categories and assigned a numerical code based upon a pre-designated range of numbers allocated to each category. Secondly, all codes were re-examined and adjusted to limit separate listings of those synonymous in meaning. The third step consisted of collapsing codes into a manageable number of "meaningful" categories for statistical analysis (Appendix 15). Concepts discriminating categories primarily emanated from the ATS acceptability criteria (ATS, 1994). Consultation with a pulmonary physiologist further confirmed the appropriateness of the categories. Lastly, the process of recoding comments was validated by a series of raters who were each asked to independently code an identical, random set of comments using final categories. From these results, percentage of agreement values were calculated (Table 2.7).

Table 2.7 Percentage of Agreement Values For Different Raters

	Rater 1 Research Assistant	Rater 2 Respiratory Research Project Manager	Rater 3 Respiratory Therapist	Rater 4 Respiratory Epidemiologist
*Original Coder	81.5%	76.9%	83.1%	87.7%
Mean Percentage of Agreement = 82.3%				

*Researcher who devised the applied coding strategy

2.11.2 Data Entry

Questionnaire data, spirometric results, and interpretation forms corresponding to each participant were entered directly into SPSS 9.0 for coding and statistical manipulations. A separate database was created for each data set (i.e. the Primary Data Set and the Secondary Data Set). However, utilization of a similar structure permitted subsequent merging of the two data sets.

Data cleaning and verification were performed on both data sets. Exploratory descriptive statistics were first used to identify obvious outliers. Each database record was then compared with its respective questionnaire. A manual review of all spirometric results was also completed. To ensure the accuracy of information transposed from interpretation forms, a random sample of 100 forms per rater was selected and checked against the appropriate database field. No data entry errors were detected.

All completed data collection forms are now stored in a secure area of a locked office in the Alberta Asthma Centre in the University of Alberta Hospitals.

2.12 Statistical Analysis

All of the statistical analyses were performed using standard statistical software packages (EpiInfo 6.0, SPSS 9.0, and LogXact 2.1).

Because the principal purpose of this study was to investigate the extent to which raters with various degrees of respiratory expertise agreed/disagreed (inter-rater reliability) in their interpretation of spirogram acceptability, initial analysis utilized Cohen's kappa, a statistic which measures level of agreement beyond that expected by chance alone[1]. Through construction of a series of two-by-two contingency tables for all possible comparisons within and between rater categories, unweighted [2] kappa coefficients were computed and tabulated. "Crude" values included raters' assessments of spirograms from all participants (i.e., in each data set). To determine whether inter-rater reliability varied according to expertise level of the technician who administered the spirometric test, spirograms were classified into technician expertise-based subsets prior to calculating coefficients. Subsequent comparative analysis employed these "stratified" kappa coefficients. For detection of statistically significant differences in rater concordance (κ), standard errors and, in turn, 95% confidence intervals[3] were calculated (Donner and Eliasziw, 1992).

Further investigation of the relationship between agreement and raters' levels of expertise involved factor analysis. Subsequent to the construction of a series of kappa matrices, Principal Components methods were applied to estimate factor loadings (rationale is discussed in Chapter 5). Plots of extracted factors (components) identified clustering of raters. Therefore, inspection of graphic results provided a basic, visual description of patterns among raters' concordance levels.

Characteristics of participants producing unacceptable spirograms were also examined.

[1] The data were checked to ensure it met critical assumptions for the kappa coefficient. These assumptions include: 1) The nominally scaled data are paired observations of the of the same phenomena 2) Ratings are assigned to categories that are mutually exclusive 3) The resulting agreement matrix is symmetric.

[2] Where there are more than two, ordered rating categories, weights may be assigned to disagreement according to the magnitude of the discrepancy. In this case, raters were asked to judge spirograms as either acceptable or not acceptable. Therefore, calculation of weighted kappas was not necessary.

[3] 95% Confidence Interval refers to the computed interval with a 95 % probability that the true value of a variable is contained within the interval (Last, 1995)

Using pre-defined criteria, spirometric results from all participants were classified as either acceptable or unacceptable (refer to Chapter 8). Assessment of relationships between the outcome variable (i.e., spirogram acceptability) and each nominal, participant variable utilized chi-square (χ^2) significance testing. Student's t-test was applied for comparison of continuous variables' means. In addition, linearity was graphically examined for all continuous variables. Those deemed to be non-linear were converted into categorical variables once appropriate categories were created.

To detect significant differences in the acceptability of spirograms associated with potential predisposing (risk) factors, odds ratios with 95% confidence intervals were calculated. When expected counts of less than five appeared in any cell, exact methods replaced chi-square tests. Importantly, prior to completion of multivariate analyses, this preliminary, bivariate analysis served as an initial screen for significant variables.

Logistic regression was used to determine participant factors associated with the spirogram outcome variable (i.e., acceptability) while adjusting for confounding factors. Due to the small, unbalanced nature of each data set (refer to Chapter 8 for a detailed discussion) exact techniques based on "conditional exact inferences" (as opposed to standard asymptotic inferences) were applied. Both forward stepwise selection and backward stepwise elimination techniques were performed. Forward stepwise selection involved sequential entry of variables into the model according to the significance of the exact (conditional scores) test (Mehta and Patel, 1999). In contrast, backward stepwise elimination required removal of statistically non-significant variables at each step of the model building process. Covariates and associated coefficients from the two models were compared in order to derive the final model. The final regression model's goodness of fit was evaluated using the Pearson chi-square test statistic and the Hosmer-Lemeshow test statistic (Hosmer and Lemeshow, 1989).

2.13 Ethical Considerations

This study received ethics approval from the University of Alberta Health Sciences Faculties, Capital Health Authority, and Caritas Health Group Health Research Ethics Board B. In accordance with their approval, all participants signed informed consent forms prior to study entry. By signing these consent forms, participants comprising each data set's study group agreed to complete questionnaires and undergo spirometric testing; interpreters evaluating spirogram acceptability agreed to rate two sets of spirogram interpretation forms. All compiled information was kept strictly confidential. Further, it was not available to anyone other than researchers directly involved in the study.

CHAPTER THREE

RESULTS - CHARACTERISTICS OF THE STUDY POPULATIONS

This chapter presents a descriptive analysis of the study populations comprising the Primary Data Set (PDS) and the Secondary Data Set (SDS). Characteristics potentially relevant to the acceptability of spirometric results are examined for each population separately. A comparison of the distribution of these characteristics between data sets is also provided.

3.1 Study Participation

Assembly of participants into subsets was based entirely on the expertise level of the technician who conducted each applicable spirometric test[1]. Factors, other than technician expertise, that may also influence spirogram acceptability were not considered. Therefore, a comparative, descriptive analysis was performed to determine the distribution of these factors across population subsets [i.e., the Minimally Trained Technician Subset (MTTS) and the Highly Trained Technician Subset (HTTS)] .

Regardless of data set, only participants who completed both the questionnaire and spirometric components were included in subsequent analyses.

3.2 Characteristics of the Study Population in the Primary Data Set

Participants in each subset (n = 50) were examined according to demographic characteristics, cardiopulmonary health history, and previous exposure to spirometry.

Demographic characteristics of the study populations are presented in Table 3.1. Results indicated a statistically significant difference between the mean ages of the two subsets (MTTS: 38.8 years, HTTS: 45.2 years; p-value = 0.04). Although not statistically

[1] Applicable spirometric tests referred to participants' initial spirometry sessions in the Primary Data Set and participants' single spirometry sessions in the Secondary Data Set.

significant, a higher proportion of participants in the HTTS was over the age of 49 years (MTTS: 27.7%, HTTS: 45.3%, p-value = 0.068). Independent of age, each subset included a gender distribution of approximately half male and half female (male - MTTS: 48.9 %, HTTS: 50.9%, female - MTTS: 51.1 %, HTTS: 49.1%). Likewise, the proportion of non-smokers was roughly equal for both subsets (MTTS: 59.6 %, HTTS: 56.6 %). In contrast, the MTTS contained a smaller percentage of participants who had smoked in the past than did the HTTS (MTTS: 21.3 %, HTTS: 30.2 %). The remaining one-fifth of the MTTS (19.1 %) and slightly less than one-sixth of the HTTS (13.2 %) were current smokers. Calculation of the average number of pack-years [1], a standardized measure of assessing smoking quantity among previous and current smokers , produced mean values of 15.3 pack-years and 18.9 pack-years for the MTTS and HTTS, respectively. Thus, with the exception of average age, none of the differences noted between population subsets were statistically significant (i.e., all p-values > 0.05).

In addition to demographic characteristics, Table 3.1 describes the cardiopulmonary health of each study population. Over half (56.6 %) of participants in the HTTS reported a history of shortness of breath, whereas, in the MTTS, a considerably lower value was observed (34.0 %). To determine severity, participants indicated the activity during which they experienced breathlessness. A similar proportion of each study population (MTTS: 10.6%, HTTS: 13.2%) recalled feeling short of breath while at rest (indicating a high degree of severity). Regardless of subset, the prevalence of physician-diagnosed asthma or wheeze was approximately 30.0 % (MTTS: 29.8.%, HTTS: 35.8%). With respect to painful breathing and frequent cough, an overwhelming majority of both populations expressed no history of either condition (no painful breathing - MTTS: 87.2 %, HTTS: 84.9 % and no frequent cough - MTTS: 78.7 %, HTTS: 67.9 %). Further, a

[1] The number of years and the quantity of cigarettes an individual smoked per day over those years can be converted into an estimated equivalent number of pack-years. Pack-years are defined as the number of years in which that same individual would have smoked one pack (20 cigarettes) per day.
The following equation for determining pack-years was applied:

$$\text{Pack-years} = \frac{\text{\# of cigarettes smoked}}{\text{day}} \times \frac{\text{one pack}}{\text{20 cigarettes}} \times \text{\# of years smoked}$$

small percentage of participants (MTTS: 10.6%, HTTS: 3.8%) had produced an abnormal chest x-ray and, among subsets combined, only one incident of both lung surgery and heart disease was reported. Importantly, assessment of cardiopulmonary health-related variables between subsets revealed no statistically significant differences (i.e., p-values > 0.05).

Table 3.1 also outlines participants' previous exposures to spirometry. Slightly over three quarters of each population (MTTS: 78.7 %, HTTS: 77.3 %) had never performed spirometry in the past. Of the remaining one quarter, most reported just one spirometric test experience prior to their involvement in the present study (MTTS: 50.0 %, HTTS: 75.0 %). The proportion of participants who completed more than four sessions was small and almost identical in each subset (MTTS: 4.3 %, HTTS: 3.8 %). Once again, analysis indicated no statistically significant differences between subsets (i.e., all p-values> 0.05).

Thus, with respect to demographic, cardiopulmonary, and spirometry-related characteristics, both subsets exhibited analogous trends.

3.3 Characteristics of the Study Population in the Secondary Data Set

Unlike in the Primary Data Set, only basic demographic information was collected for participants (n = 200) in the Secondary Data Set (refer to section 2.5.1).

Table 3.2 provides a demographic comparison of the study populations forming the Secondary Data Set. The average age of the MTTS was statistically significantly higher than that of the HTTS (MTTS: 38.4 years, HTTS: 34.0 years). However, participants over 50 years of age comprised nearly equal proportions of each subset (MTTS: 11.0%, HTTS: 8.0%). Additionally, about three-fifths of both populations consisted of females (MTTS: 59.0 %, HTTS: 63.0 %). With regard to smoking status, over half of the participants in each subset had never smoked (MTTS: 55.0 %, HTTS: 78.0 %).

Compared with the HTTS, the percentage of participants in the MTTS who currently smoked (i.e., at the time of data collection) was three times higher (MTTS: 24.0 %, HTTS: 7.0 %). Similarly, a greater number of participants in the MTTS reported having smoked in the past (MTTS: 21.0 %, HTTS: 15.0 %). Nevertheless, apparent differences between subsets were not statistically significant. Additionally, no statistically significant variance in the average number of pack-years was observed (MTTS: 14.0 years, HTTS: 9.8 years).

In general, based primarily on participant age (dichotomized at 50 years), gender, and smoking status, the above findings suggested similarity of subset populations.

3.4 Comparability of the Primary and Secondary Data Sets

To identify differences in the study populations that could influence not only the interpretability, but also, the generalizability of subsequent results, a comparative analysis of demographic information common to both data sets was performed.

The results showed statistically significant discrepancies in the age distribution of the Primary and Secondary Data Sets (refer to Table 3.3). Although participants in both groups spanned analogous age ranges, the mean age of the Primary Data Set was roughly 6 years greater than that of the Secondary Data Set (PDS: 42.2, SDS: 36.2). Also, the Primary Data Set included four times more participants above 49 years of age than did the Secondary Data Set (PDS: 37.0%, SDS: 9.5%). Regarding gender, each data set included similar proportions of females (female - PDS: 50.0%, SDS: 61.0%). In addition to the mean number of pack-years, the distribution of current, past, and non-smokers in the two data sets closely resembled one another (p-value = 0.647). Therefore, with the exception of age, participants forming the Primary and Secondary Data Sets were comparable.

3.5 Conclusions

Based on the distribution of demographic characteristics recognized in the literature as affecting spirogram acceptability (i.e., age and gender), participant populations who formed the MTTS and HTTS of the Secondary Data Set did not differ statistically. Thus, it was concluded that technicians of either expertise level administered spirometry to similar populations. Conversely, in the Primary Data Set, participants comprising the HTTS were slightly older than those of the MTTS. Past studies have established an increased risk of test failure among participants over the age of 50 (refer back to Section 1.6.2). Consequently, statistical techniques (discussed in Chapter 8) were applied to control for the potential confounding effect of age on the relationship between technician expertise and test acceptability.

It was also noted that, despite their similarity in gender, the Primary Data Set and the Secondary Data Set varied with respect to the proportion of participants over the age of 49 years. The "clinical significance" of this difference is uncertain.

Table 3.1 Distribution of Participants in Each Subset of the Primary Data Set According to Demographic, Cardiopulmonary Health, and Spirometry-Related Characteristics

	Minimally Trained	Highly Trained	
	Technician Subset	Technician Subset	p-value*
Characteristic	n (%)	n (%)	
Demographic			
Age (years)			
Mean (sd)	38.8 (15.2)	45.2 (14.7)	0.035[†]
Range	18-77	19-77	
< 50 years	34 (72.3)	29 (54.7)	0.068
≥ 50 years	13 (27.7)	24 (45.3)	
Gender			
Male	23 (48.9)	27 (50.9)	>0.999
Female	24 (51.1)	26 (49.1)	
Smoking Status			
Current Smoker	9 (19.1)	7 (13.2)	0.524
Past Smoker	10 (21.3)	16 (30.2)	
Non-Smoker	28 (59.6)	30 (56.6)	
Pack-years (for current and past smokers)			
Mean (sd)	15.3 (19.3)	18.9 (21.3)	0.562[†]
Cardiopulmonary			
History of Shortness of Breath			
No	31 (66.0)	23 (43.4)	0.221
Yes	16 (34.0)	30 (56.6)	
Generalized:	2 (4.3)	4 (7.6)	
Specified:			
while climbing stairs	5 (10.6)	13 (24.5)	
while walking	4 (8.5)	6 (11.3)	
while at rest	5 (10.6)	7 (13.2)	
History of Painful Breathing			
No	41 (87.2)	45 (84.9)	0.483
Yes	6 (12.8)	8 (15.1)	
History of Wheeze or Asthma			
No	33 (70.2)	34 (64.2)	0.400
Yes	14 (29.8)	19 (35.8)	
History of Frequent Cough			
No	37 (78.7)	36 (67.9)	0.117
Yes	10 (21.3)	17 (32.1)	
without sputum	5 (10.6)	3 (5.7)	
with sputum	5 (10.6)	14 (26.4)	

Table 3.1 Continued

Characteristic	Minimally Trained Technician Subset n (%)	Highly Trained Technician Subset n (%)	p-value
Abnormal Chest X-Ray			
No	42 (89.4)	51 (96.2)	0.249
Yes	5 (10.6)	2 (3.8)	
History of Lung Surgery			
No	47 (100.0)	52 (98.1)	>0.999
Yes	0 (0.0)	1 (1.9)	
History of Heart Disease			
No	47 (100.0)	52 (98.1)	>0.999
Yes	0 (0.0)	1 (1.9)	
Previous Spirometric Tests			
Completion of Spirometry Prior to Involvement in Present Study			
No	37 (78.7)	41 (77.3)	0.663
Yes	10 (21.3)	12 (22.7)	
Number of Previous Spirometry Sessions:			
1 session	5 (10.7)	9 (17.0)	
2 sessions	1 (2.1)	0 (0.0)	
3 sessions	1 (2.1)	1 (1.9)	
4 sessions	1 (2.1)	0 (0.0)	
≥ 5 sessions	2 (4.3)	2 (3.8)	
Total Number of Participants	47	53	

* p-value based on Fisher's Exact Test statistic, unless otherwise indicated

† p-value based on t-test statistic

Table 3.2 Distribution of Participants in Each Subset of the Secondary Data Set According to Demographic Characteristics

Characteristic	Minimally Trained Technician Subset n (%)	Highly Trained Technician Subset n (%)	p-value*
Number of Participants	100 (100.0)	100 (100.0)	
Age (years)			
Mean (sd)	38.4 (10.3)	34.0 (9.4)	0.002†
Range	18-76	21-53	
< 50 years	89 (89.0)	92 (92.0)	0.469
≥ 50 years	11 (11.0)	8 (8.0)	
Gender			
Male	41 (41.0)	37 (37.0)	0.562
Female	59 (59.0)	63 (63.0)	
Smoking Status			
Current Smoker	24 (24.0)	7 (7.0)	0.180
Non-Smoker	55 (55.0)	78 (78.0)	
Past Smoker	21 (21.0)	15 (15.0)	
Pack-years (for current and past smokers)			
Mean (sd)	14.0 (14.1)	9.8 (12.4)	0.236†

* p-value based on chi-square statistic, unless otherwise indicated

† p-value based on t-test statistic

Table 3.3 Distribution of Participants in the Primary and Secondary Data Sets According to Demographic Characteristics

Characteristic	Primary Data Set n (%)	Secondary Data Set n (%)	p-value*
Age (years)			
Mean (sd)	42.17 (15.23)	36.19 (10.04)	0.001†
Range	18-77	18-76	
< 50 years	63 (63.0)	181 (90.5)	0.000
≥ 50 years	37 (37.0)	19 (9.5)	
Gender			
Male	50 (50.0)	78 (39.0)	0.069
Female	50 (50.0)	122 (61.0)	
Smoking Status			
Current Smoker	16 (16.0)	31 (15.5)	0.647
Past Smoker	28 (28.0)	36 (18.0)	
Non-Smoker	56 (56.0)	133 (66.5)	
Pack-years (for current and past smokers)			
Mean (sd)	17.1 (20.2)	12.6 (13.6)	0.709†
Total Number of Participants	100	200	

* p-value based on chi-square statistic, unless otherwise indicated

† p-value based on t-test statistic

CHAPTER FOUR

RESULTS - INTERPRETATION OF SPIROGRAM ACCEPTABILITY ACCORDING TO RATER EXPERTISE

This chapter summarizes concordance among raters at various levels of respiratory expertise who evaluated the acceptability of participants' spirometric results.

4.1 Determination of Inter-rater Reliability

One of the study's primary objectives was to determine inter-rater reliability for the acceptability of spirograms. In general, inter-rater reliability studies assess agreement between raters who assign a specific characteristic or trait to a pre-defined category. The "degree" of agreement provides an estimate of the precision of such judgements in the absence of a 'gold standard' against which to evaluate their accuracy. Therefore, variance, attributable to the rating process, is quantified (Posner et al, 1990).

4.1.1 Cohen's Kappa Statistic

The most commonly employed measure of inter-rater reliability is the kappa statistic (κ). Defined as chance-corrected concordance, kappa compares the observed level of agreement with that expected by chance alone (Last, 1995). The general expression for kappa is:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where[1]: P_o = proportion of observed agreement = $\left\{\frac{\text{number of agreements}}{\text{number of paired observations}}\right\}$

P_e = proportion of chance agreement = $\sum_{i=1}^{k} \left\{\frac{\text{(row marginal)(column marginal)}}{\text{(number of paired observations)}^2}\right\}$

let k = number of cells on the diagonal

[1] Given a square r x c contingency table, P_o is calculated by summing the number of agreements indicated on the diagonal and dividing by the total number of paired observations. To determine P_e, the row and column marginal totals for each cell on the diagonal are multiplied together and divided by the total number of observations. Division of this expected frequency by the total number of observations yields P_e. These proportions are summed across all of the cells on the diagonal to obtain the total proportion of expected agreement.

Kappa coefficients, in theory, can range from -1 to +1. A negative value indicates that measurements agree less often than expected by chance while a positive value demonstrates agreement more often than expected by chance. If concordance is complete (i.e., perfect agreement) kappa achieves its maximum value of +1. A value of zero represents no agreement beyond that which could be attributed to chance alone.

4.1.2 Calculation of Unweighted Kappa Coefficients

For this study, kappa coefficients corresponding to all rater comparisons across respiratory expertise-based categories were obtained, tabulated and examined. In particular, analysis involved calculation of unweighted kappa values. Inter-rater reliability studies requiring assignment of a characteristic to one of several, ordered categories frequently weight disagreement between raters according to the magnitude of the discrepancy [e.g., Observations closer to the diagonal (on a standard contingency table) are less discrepant and, thus, considered less serious than those farther away]. The resulting coefficient is termed a weighted kappa value (Altman, 1991). In contrast, the present study employed a dichotomous classification scheme. Raters evaluated spirogram quality as either acceptable or unacceptable. Because the degree of disagreement between raters remained constant, weighting of kappa coefficients was not appropriate.

4.1.3 Assessment of Kappa Coefficients

Assessment of each kappa coefficient included testing its statistical significance and evaluating its relative magnitude.

i) Statistical Significance of Kappa

To identify whether raters agreed significantly more or less than expected by chance, alone (i.e., $\kappa \neq 0$), standard errors and, in turn, 95% confidence intervals were calculated. The absence of zero values from such intervals indicated statistical significance of corresponding kappa coefficients. To detect statistically significant differences between kappa scores, confidence intervals were examined for complete separation (i.e., no overlapping values). The procedure selected for constructing confidence intervals provides more accurate coverage levels in samples of smaller size than is recognized for alternative methods (Kraemer and Bloch, 1989). The following equations were applied:

To calculate standard error for kappa:

$$se(\kappa) = \left\{\frac{1-\kappa}{N}\left[(1-\kappa)(1-2\kappa) + \frac{\kappa(2-\kappa)}{2\pi(1-\pi)}\right]\right\}^{1/2}$$

where: N = the number of subjects

$$\pi = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2} X_{ij}$$

Let X_{ij} denote the rating for the ith subject assigned by the jth rater

To calculate 95% confidence intervals about kappa:

$$95\ \%\text{C.I.}(\kappa) = \kappa \pm 1.96\ se(\kappa)$$

(Donner and Eliasziw, 1992)

ii) Interpretation of Kappa Coefficients

Interpretation of each kappa value's relative magnitude was based on suggestions proposed by Landis and Koch (Landis and Koch, 1977). Although arbitrary in nature, such recommendations have become incorporated into the literature as standard criteria (refer to Table 2.5). Thus, with respect to strength of agreement, kappa values of less than 0.00, 0.00 to 0.20, 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80 and 0.81 to 1.00 denoted poor, slight, fair, moderate, substantial, and almost perfect agreement, respectively (Landis and Koch, 1977).

4.1.4 <u>Summarizing Rater Agreement Within and Between Categories</u>

To summarize rater agreement at the category level and, in turn, simplify comparative analyses, the arithmetic mean value of kappa coefficients calculated from pair-wise combinations of individual raters within or between categories was computed [e.g., kappa values for rater combinations 7 and 9, 7 and 10, 8 and 9, and 8 and 10 were averaged to yield a single kappa value representing the level of agreement between Respiratory Epidemiologists (raters 7 and 8) and Pulmonary Specialists (raters 9 and 10)]. A comparison of these results to those collected using an 'average of ranks' approach served to verify the appropriateness of this technique (Appendix 16). Both methods produced similar findings, supporting the decision to present only arithmetic mean-derived kappa values. These values were then interpreted according to methods specified for interpreting individual kappa scores (Section 4.1.3.ii).

4.2 Results of Agreement Between Raters

4.2.1 Introduction

As mentioned previously, each participant contributed a pair of spirograms to either the Primary or Secondary Data Set. All pairs included one best test[1] spirogram (characterized by its production of the largest sum of FEV_1 plus FVC) and one complementary spirogram (which displayed the second largest sum of FEV_1 plus FVC). For each spirogram type, kappa coefficients, both crude and stratified, were calculated. Crude (i.e., unstratified) kappa coefficients incorporated data from all participants' either best test spirograms or complementary spirograms, providing a general indication of the degree of inter-rater reliability within and between respiratory expertise categories. An additional objective of this study involved determining the extent to which the expertise level of the test administrator (i.e., spirometry technician) influenced concordance between raters. Recall that participants in each data set were assembled into subsets according to the training and experience of technicians who conducted respective spirometric tests. Subsequent calculation of kappa coefficients for separate subsets produced a series of stratified values[2]. These values were then compared to assess the effect of technician expertise on rater concordance.

[1] According to ATS criteria, best test spirograms are those first evaluated as acceptable which produced the largest sum of FEV_1 plus FVC (ATS, 1994).

[2] Stratification of kappa values was based on expertise level of pulmonary function technicians.

4.2.2 Agreement Between Raters With Similar Respiratory Expertise (Crude Kappa Results)

Inter-rater reliability regarding the acceptability of spirometric results was first examined between raters whose levels of respiratory expertise were similar.

i) The Primary Data Set

Within-category agreement between raters for the acceptability of participants' best test spirograms (from first spirometry sessions) is presented in Table 4.1. Among the three, Non-certified, Minimally Trained, Respiratory Research Assistants, kappa values, ranging from 0.21 to 0.52, indicated slight to moderate agreement. In addition, all respective confidence intervals overlapped, suggesting that differences between coefficients were not statistically significant. However, the "conservative"[1] approach used for constructing upper and lower limits yields wider intervals than do alternative methods (Walter, 1999). Therefore, observation of values that marginally overlapped reflected "clinically significant" discrepancies in agreement between individual raters of a common, minimal, respiratory expertise level. Within the Certified, Respiratory Technician category, inter-rater comparisons produced a narrower range of kappa coefficients indicating fair to moderate agreement (0.29 to 0.51). The extent to which corresponding confidence intervals overlapped was greater than that observed for the Non-certified, Minimally Trained, Respiratory Research Assistants. Within the Respiratory Epidemiologist category, a single kappa value of 0.58 signified moderate inter-rater reliability. Analogous results were reported for the two pulmonary specialists (kappa = 0.50). Thus, regardless of expertise, concordance between raters whose respiratory training and experience were considered equivalent did not exceed moderate levels. The absence of clear homogeneity in agreement across pairs of raters within any single category suggested that the degree of inter-rater reliability was not merely a reflection of raters' respiratory expertise.

[1] "Conservative" methods error on the side of safety (i.e., statistical non-significance).

Discrepant kappa values, calculated subsequent to addition of one or more raters to each of the Respiratory Epidemiologist and Pulmonary Specialist categories, would aid in confirming the above conclusion.

Results for participants' complementary spirograms were similar to those outlined for participants' best test spirograms (refer to Table 4.2). Among pairs of Non-certified, Minimally Trained, Respiratory Research Assistants, kappa values spanning from 0.18 to 0.45 demonstrated slight to moderate agreement. The Certified, Respiratory Technician group generated both a narrower range of kappa coefficients (0.35 to 0.50) and a set of corresponding confidence intervals that significantly overlapped. Inter-rater reliability within the Respiratory Epidemiologists and Pulmonary Specialists categories produced a similar pattern of almost identical kappa values (0.45 and 0.50, respectively). Thus, no well-defined variances in the degree of inter-rater reliability for participants' best test spirograms or their complementary counterparts were detected.

ii) The Secondary Data Set

Within-category rater concordance for the acceptability of best test spirograms in the Secondary Data Set is summarized in Table 4.3. Although kappa values obtained from comparisons between Non-certified, Minimally Trained, Respiratory Research Assistants reflected a higher level of agreement (0.45 to 0.70) than did those in the Primary Data Set (Primary Data Set: slight to moderate agreement; Secondary Data Set: moderate to substantial agreement), their confidence intervals overlapped only marginally. Among Certified Respiratory Technicians, kappa coefficients of 0.21 to 0.51, indicating fair to moderate inter-rater reliability, coincided with those of the Primary Data Set. Despite calculation of slightly lower kappa values, moderate levels of agreement were, once again, achieved by both Respiratory Epidemiologists and Pulmonary Specialists (0.48 and 0.40, respectively). Thus, with the exception of the Non-certified, Minimally Trained, Respiratory Research Assistant group, the Primary and Secondary Data Sets presented parallel findings.

Table 4.4 describes rater agreement for the acceptability of complementary spirograms in the Secondary Data Set. In two of the four expertise-based categories, concordance levels were one "strength of agreement" interval lower than those reported for best test spirograms. Specifically, kappa values derived from the Non-certified, Minimally Trained, Respiratory Research Assistants group spanned from 0.33 to 0.61, and, in turn, suggested only fair to moderate inter-rater reliability. With respect to the Respiratory Epidemiologist group, a kappa value of 0.28 also indicated a comparative reduction in consensus. Levels in each of the Certified, Respiratory Technician and Pulmonary Specialist categories remained constant across spirogram types (0.23 to 0.48 and 0.45, respectively). Therefore, contrary to the Primary Data Set, results pertaining to complementary spirograms did not completely emulate those for best test spirograms. Moreover, no correlation between differences in agreement across spirogram types and raters' respiratory expertise was noted.

Because the two data sets generated partially conflicting results, not only the stability, but, also, the significance of apparent differences in inter-rater reliability between complementary spirograms and best test spirograms in the Secondary Data Set was uncertain.

iii) Summary Analysis of Agreement Between Raters With Similar Respiratory Expertise (Average Kappa Values)

To further clarify the relationship between within-category agreement and raters' respiratory expertise, average kappa values, which generalized agreement within categories, were compared. Conclusions replicated those derived from inspection of multiple kappa scores at respective respiratory expertise levels.

Table 4.5 presents results of the Primary Data Set. For best test and complementary spirograms, inter-rater reliability among Non-certified, Minimally Trained, Respiratory Research Assistants was fair. All other categories generated summary kappa values of moderate strength. Thus, no variation in agreement across spirogram types was found. Based on these results, there appeared to be a relationship between higher "overall" inter-rater reliability and professional respiratory expertise. However, the comparatively low mean value for the Non-certified, Minimally Trained, Respiratory Research Assistants category was calculated from a set of notably discrepant individual kappa coefficients. Average scores do not reflect such discrepancies. Consequently, this inferred relationship may be unreliable.

Results of the Secondary Data Set are outlined in Table 4.6. Average kappa values corresponding to participants' best test spirograms reflected moderate concordance within groups of Non-certified, Minimally Trained, Respiratory Research Assistants, Respiratory Epidemiologists, and Pulmonary Specialists. In the Certified, Respiratory Technicians Category, slightly lower values indicated fair inter-rater reliability. Complementary spirograms, with the exception of the Respiratory Epidemiologists group, received equivalent degrees of concordance to their best test counterparts. Consequently, neither a relationship between the degree of within-category agreement and raters' respiratory expertise nor any substantial variation in levels across spirogram types was detected.

4.2.3 Effect of Spirometry Technician Expertise on Agreement Between Raters With Similar Respiratory Expertise (Stratified Kappa Results)

i) The Primary Data Set

Rater agreement for the acceptability of best test spirograms from sessions directed by minimally trained and highly trained technicians is summarized in Tables 4.7 and 4.8, respectively. Within categories of more than two raters, differences[1] in inter-rater reliability between technician expertise levels varied among pairwise combinations of

[1] Kappa values between technician expertise levels were deemed statistically significantly different if: 1)at most, a marginal degree of overlap in confidence intervals was noted and 2)"strength of agreement" changed by at least one interval (Landis and Koch, 1977). In the present study any cited difference met these criteria.

raters. Although one of three pairs comprising the Non-certified, Minimally Trained, Respiratory Research Assistants Category exhibited significantly more agreement (i.e., an increase of at least one "strength of agreement" interval) for tracings produced by the minimally trained technician, the other two reported greater levels for spirograms collected from the highly trained technician. Similarly, in one of the pairs consisting of Certified, Respiratory Technicians, stronger concordance corresponded to the minimally trained technician's spirograms; whereas among the remaining two-thirds, agreement remained constant across technician expertise levels. Between Respiratory Epidemiologists, equivalent degrees of inter-rater reliability were also exhibited. However, spirograms from the highly trained technician generated greater concordance in the Pulmonary Specialists Category. Importantly, because these latter two categories each included only one pair of raters, the generalizability of their results could not be assessed. Based exclusively on these findings, independent of raters' expertise, no definite, positive or negative correlation between strength of agreement and technician expertise existed.

Similar conclusions were derived from complementary spirograms (refer to Tables 4.9 and 4.10). Prominent "technician effects" appeared to be restricted to categories representing the two extremes in respiratory expertise (i.e., the Non-certified, Minimally Trained, Respiratory Research Assistants and Pulmonary Specialists). Apart from those comprising the Non-certified, Minimally Trained, Respiratory Research Assistants Category, no pair demonstrated greater agreement for tracings provided by the highly trained technician. Among two-thirds of Certified, Respiratory Technicians, inter-rater reliability did not vary between technician expertise levels. This trend was also observed in the Respiratory Epidemiologist Category. With respect to both Pulmonary Specialists and the remaining pair of Certified Respiratory Technicians, agreement significantly decreased with technician expertise. Therefore, in accordance with findings from participants's best test spirograms, the effect of technician expertise level on rater agreement was both pair and, to an extent, category-specific. Further, the absence of uniform, within-category results across best test and complementary spirograms, suggested that such "technician effects" were, in addition, spirogram specific.

ii) The Secondary Data Set

Inter-rater reliability for participants' best test spirograms was examined according to technician expertise in Tables 4.11 and 4.12. Unlike those of the Primary Data Set, results, in general, did not indicate the presence of a "technician effect" on raters' levels of concordance. Throughout all categories, combined, only one pair, comprised of Non-Certified, Minimally Trained, Respiratory Research Assistants, generated significantly greater agreement for the highly trained technician's spirograms. Among Pulmonary Specialists, Respiratory Epidemiologists and the majority of both Non-certified, Minimally Trained, Respiratory Research Assistants and Certified Respiratory Technicians, spirograms from tests conducted by either minimally trained technicians or highly trained technicians produced equivalent degrees of concordance (i.e., no difference in agreement between technician expertise levels). Nevertheless, in each category consisting of more than two raters, results of one pair did not coincide with this general pattern. Therefore, similar to the Primary Data Set, apparent "technician effects" were not completely consistent across raters with similar respiratory expertise.

Results from complementary spirograms are presented in Tables 4.13 and 4.14. Resembling the Primary Data Set, concordance levels in all three pairs of Non-certified, Minimally Trained, Respiratory Research Assistants significantly increased with technician expertise. However, no other category demonstrated distinguishable differences in the degree of agreement between technicians' spirograms. Consequently, a "technician effect", for complementary spirograms, was not detected among pairs of raters with professional respiratory expertise (i.e., Certified Respiratory Technicians, Respiratory Epidemiologists and Pulmonary Specialists).

iii) Summary Analysis of the Effect of Technician Expertise on Agreement Between Raters With Similar Respiratory Expertise

The capacity of average kappa values to contribute a "meaningful" summary of the effect of technician expertise on agreement between raters with similar expertise is questionable. First, "technician effects" appeared to be discrepant within categories comprised of more than two raters (Note: These were the only categories in which average kappa scores differed from individual scores). Comparison of average kappa values failed to capture this information. Second, because a method for calculating confidence intervals around average kappa values could not be determined, criteria used to assess the statistical significance of differences between kappa values in preceding analyses were no longer applicable. To both acknowledge and compensate for these limitations, inferences derived from visual inspection of mean scores that diverged from previously stated findings were regarded as less accurate. In the absence of confidence intervals, average differences in kappa scores were considered to be significant if their magnitudes were equal to or greater than 0.2, the width of each "strength of agreement interval" (Landis and Koch, 1977).

Average kappa scores in the Primary Data Set are tabulated in Table 4.15. For best test spirograms, only the Pulmonary Specialists category exhibited a "technician effect". Specifically, tracings from the highly trained technician generated stronger concordance. With respect to results of complementary spirograms, the Non-certified, Minimally Trained, Respiratory Research Assistants category exhibited greater levels of agreement for those of the highly trained technician while the reverse finding was observed between Pulmonary Specialists. Table 4.16 presents mean scores in the Secondary Data Set. Regardless of rater expertise, no "technician effects" were observed for best test spirograms. With the exception of the Non-certified, Minimally Trained, Respiratory Research Assistants category, in which agreement appeared to increase with technician expertise, identical results were presented for complementary spirograms. Thus, the effect of technician expertise on agreement between raters with similar respiratory expertise was not only expertise-specific but, also, only spirogram type-specific.

4.2.4 Agreement Between Raters With Differing Respiratory Expertise (Crude Kappa Results)

Inter-rater reliability for the acceptability of participants' spirometric results was also examined between raters whose levels of respiratory expertise were diverse. Adhering to the rationale discussed above (Section 4.2.2), some differences between kappa values (representing internal concordance and those representing concordance between categories) were considered potentially important despite their statistical non-significance if associated confidence intervals overlapped only marginally.

i) The Primary Data Set

a) Comparisons Between Non-certified, Minimally Trained, Respiratory Research Assistants and Raters With Professional Respiratory Expertise

With respect to the acceptability of best test spirograms, variations in agreement levels between Non-certified, Minimally Trained, Respiratory Research Assistants and raters with professional respiratory expertise were discrepant across combinations of raters (refer to Table 4.1). Pairs, each comprised of one Non-certified, Minimally Trained, Respiratory Research Assistant and one Certified, Respiratory Technician, generated kappa values ranging from 0.01 to 0.59 and, subsequently, reflected slight to moderate agreement. Although concordance reached a maximum level comparable to that observed within each of these categories (0.52 among Non-certified, Minimally Trained, Respiratory Research Assistants; 0.51 among Certified, Respiratory Technicians), neither category, internally, produced a minimum value indicative of only slight agreement. Therefore, pair-specific results suggested that rater concordance between the two expertise levels was either equivalent to, or slightly less than that within expertise levels. Combinations of Non-certified, Minimally Trained, Respiratory Research Assistants and Respiratory Epidemiologists exhibited kappa values of 0.22 to 0.58, a range almost identical to that of pairs within categories consisting exclusively of Non-certified, Minimally Trained, Respiratory Research Assistants or Respiratory Epidemiologists.

However, one set of individual kappa scores[1] corresponding to a minimally trained rater indicated significantly greater agreement between rather than within these categories. Among Non-certified, Minimally Trained, Respiratory Research Assistants and Pulmonary Specialists, results were also pair-specific (i.e., kappa values were not consistent across different pairwise combinations of raters). While the majority of those from within and between these two categories demonstrated similar degrees of inter-rater reliability, certain heterogeneous pairs representing combined expertise levels displayed significantly less agreement. Thus, concordance did not systematically vary between raters with minimal, respiratory expertise and those with professional, respiratory expertise. For complementary spirograms, kappa values comparable to those noted for best test spirograms revealed analogous, pair-specific trends (refer to Table 4.2).

b) Comparisons Between Raters With Professional Respiratory Expertise

Among pairs which combined Certified, Respiratory Technicians with either Respiratory Epidemiologists or Pulmonary Specialists, kappa values spanned an analogous set of values (0.26 to 0.63 and 0.31 to 0.61, respectively). Similarly, pairs of Respiratory Epidemiologists and Pulmonary Specialists produced coefficients ranging from 0.37 to 0.63. Thus, fair to substantial concordance was exhibited between all categories. Inspection of consecutive kappa scores (presented in rows and columns on each table) reflected statistically non-significant differences in agreement between raters across professional respiratory expertise levels. Unlike those involving Non-certified, Minimally Trained, Respiratory Research Assistants, no pair generated a kappa value signifying only slight concordance. Similarly, for complementary spirograms, coefficients indicating "slight" inter-rater reliability were restricted to pairs of Non-certified, Minimally Trained Respiratory Research Assistants and Pulmonary Specialists (refer to Table 4.2). Therefore, irrespective of spirogram type, greater agreement was demonstrated between raters with professional, respiratory expertise.

[1] Through multiple, pairwise comparisons of each rater with all other raters (e.g., pairing rater 1 separately with raters 2-10), sets of individual kappa scores were assembled.

ii) The Secondary Data Set

a) Comparisons Between Non-certified, Minimally Trained, Respiratory Research Assistants and Raters With Professional Respiratory Expertise

Contrary to the Primary Data Set, multiple, pairwise combinations of raters produced consistent findings (refer to Table 4.3). All pairs involving Non-certified, Minimally Trained, Respiratory Research Assistants and raters with professional, respiratory expertise (i.e., Certified, Respiratory Technicians, Respiratory Epidemiologists, or Pulmonary Specialists) exhibited statistically significantly lower concordance than did those containing exclusively Non-certified, Minimally Trained, Respiratory Research Assistants. Comparisons of minimally trained raters with Certified, Respiratory Technicians, Respiratory Epidemiologists, and Pulmonary Specialists produced overlapping sets of kappa coefficients that ranged from 0.03 to 0.54, 0.26 to 0.46, and 0.24 to 0.40, respectively. Although both maximum and minimum kappa coefficients corresponded to pairs of Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians, in general, certain values were not characteristic of particular expertise category combinations. Further, discrepancies in pair specific agreement appeared to decrease as the difference in expertise between raters increased. Pairs contrasting the two extremes in respiratory expertise (i.e., Non-certified, Minimally Trained, Respiratory Research Assistants and Pulmonary Specialists) produced the narrowest range of kappa scores. Despite calculation of slightly lower kappa values, parallel trends were established for participants' complementary spirograms (refer to Table 4.4). Again, concordance levels of pairs involving Non-certified, Minimally Trained, Respiratory Research Assistants and raters with professional respiratory expertise were consistently lower than those of pairs comprised exclusively of Non-certified, Minimally Trained, Respiratory Research Assistants. Collectively, these findings suggested that strength of concordance was inversed related to similarity in rater expertise level.

b) Comparisons Between Raters With Professional Respiratory Expertise

No uniform effect of varying raters' levels of professional respiratory expertise on concordance was detected (refer to Table 4.3). Between Certified, Respiratory Technicians and both Respiratory Epidemiologists and Pulmonary Specialists, kappa values indicated slight to moderate agreement. Additionally, differences in "serial" kappa values, which traced concordance achieved by pairs of raters across professional respiratory expertise categories, were statistically non-significant. Similarly, comparisons within and between the Respiratory Epidemiologists and Pulmonary Specialists categories yielded equivalent degrees of agreement. In contrast, kappa scores from complementary spirograms reflected variances among expertise levels (refer to Table 4.4). Agreement between Certified, Respiratory Technicians and one Respiratory Epidemiologist reached levels of statistically lower significance than observed within the Certified, Respiratory Technician Category. While one Certified, Respiratory Technician, when paired with either Pulmonary Specialist exhibited a similar trend, the majority generated values similar to those derived internally. Across Respiratory Epidemiologists and Pulmonary Specialists, differences in concordance were also inconsistent. Half of these pairs displayed significantly greater concordance than noted between Respiratory Epidemiologists. Therefore, based on pair-specific results for complementary spirograms, a coherent effect of professional, respiratory expertise level on inter-rater reliability was not established.

iii) *Summary Analysis of Agreement Between Raters With Different Levels of Respiratory Expertise*

Average kappa values were examined in order to verify conclusions inferred from pair-specific scores.

a) Comparisons Between Non-certified, Minimally Trained, Respiratory Research Assistants and Raters With Professional Respiratory Expertise

In the Primary Data Set, all mean kappa coefficients corresponding to participants' best test spirograms indicated fair concordance (refer to Table 4.5). For participants' complementary spirograms, agreement among pairs of Non-certified, Minimally Trained, Respiratory Research Assistants and of those comparing Non-certified, Minimally

Trained, Respiratory Research Assistants with either Certified, Respiratory Technicians or Respiratory Epidemiologists was also fair. Observation of only "slight" concordance across pairs of Non-certified, Minimally Trained, Respiratory Research Assistants and Pulmonary Specialists reflected discrepancies in pair-specific values. Thus, based on average kappa scores, no definite, systematic variation in agreement between raters with minimal respiratory expertise and those with different professional levels of respiratory expertise was found.

For best test spirograms in the Secondary Data Set, mean kappa coefficients also indicated "fair" agreement between categories (refer to Table 4.6). However, average concordance within the Non-certified, Minimally Trained, Respiratory Research Assistants Category was "moderate". Therefore, raters with similar, minimal respiratory expertise appeared to agree more strongly among themselves than with those at a professional level. Parallel analysis of complementary spirograms contributed identical findings (Refer to Table 4.6).

b) Comparisons Between Raters With Professional Respiratory Expertise

According to average kappa values for best test spirograms in the Primary Data Set, Respiratory Epidemiologists and Pulmonary Specialists exhibited greater agreement ("moderate") than did either expertise level on individual comparison with Certified, Respiratory Technicians (refer to Table 4.5). However, participants' complementary spirograms generated equal concordance ("moderate") across all professional respiratory-expertise based comparisons.

Results congruent with those of the Primary Data Set were exhibited for participants' best test spirograms in the Secondary Data Set (refer to Table 4.6). Again, agreement between Respiratory Epidemiologists and Pulmonary Specialists was "moderate". Comparison of these groups with those consisting of Certified, Respiratory Technicians resulted in lower ("fair") concordance. In contrast, participants' complementary spirograms generated "fair" agreement between Pulmonary Specialists and both Certified Respiratory

Technicians and Respiratory Epidemiologists and only "slight" levels among Certified, Respiratory Technicians and Respiratory Epidemiologists. As discussed previously, average values did not reflect discrepancies in kappa coefficients which corresponded to pairs of Certified, Respiratory Technicians and Respiratory Epidemiologists. Therefore, this finding might inaccurately depict differences affiliated with raters' level of respiratory expertise.

4.2.5 Effect of Spirometry Technician Expertise on Agreement Between Raters With Different Respiratory Expertise (Stratified Kappa Results)

i) The Primary Data Set

a) Comparisons Between Non-Certified, Minimally Trained, Respiratory Research Assistants and Raters With Professional Respiratory Expertise

Concordance regarding best test spirograms from sessions directed by minimally trained or highly technicians varied across pairs of raters representing synonymous expertise levels (refer to Tables 4.7 and Tables 4.8). Between Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians, kappa coefficients of 0.02 to 0.66 for the minimally trained technician's spirograms indicated slight to substantial agreement. With respect to the highly trained technician's spirograms, observed values from 0.06 to 0.46 signified slight to moderate concordance. Similarly, in pairs consisting of Non-certified, Minimally Trained, Respiratory Research Assistants and either Respiratory Epidemiologists or Pulmonary Specialists, kappa scores for the highly trained technician's tracings spanned a narrower range of values than did those for the minimally trained technician (minimally trained technician's spirograms: 0.06 to 0.56 and -0.06 to 0.34, respectively; highly trained technician's spirograms: 0.35 to 0.59 and 0.23 to 0.44, respectively). Therefore, consistency in pair-specific agreement appeared to increase with technician expertise. Scores corresponding to identical raters at each technician expertise level were also compared. Two-thirds of pairs combining Non-certified, Minimally Trained, Respiratory Research Assistants with Certified, Respiratory Technicians and Respiratory Epidemiologists and half of those incorporating Pulmonary Specialists exhibited no statistically significant difference in agreement between technician's tracings. Among remaining pairs, apparent "technician effects" differed across professional respiratory expertise categories. Between Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians, stronger concordance corresponded to spirograms from the minimally trained technician. Conversely, parallel comparisons involving Respiratory Epidemiologists and Pulmonary Specialists generated greater levels of agreement for the highly trained technician's spirograms. Thus, based on results from remaining pairs, concordance between raters who exemplified the extremes in respiratory backgrounds strengthened with technician expertise. Similar trends were determined utilizing complementary spirograms (refer to Tables 4.9 and 4.10). Once again, the majority of pairs produced equivalent degrees of agreement for each technician's spirograms. However, across all other pairs, regardless of their professional respiratory expertise category, increased agreement coincided with the highly trained technician's tracings. Thus, the existence of a positive correlation between strength of concordance and technician expertise was confined to specific pairs of raters.

b) Comparisons Between Raters With Professional Respiratory Expertise

"Technician effects" on concordance between raters in different professional respiratory expertise categories were not uniform (refer to Tables 4.7 and 4.8). At least two-thirds of pairs, comprised of either Certified, Respiratory Technicians and Respiratory Epidemiologists or Pulmonary Specialists and Respiratory Epidemiologists demonstrated equal concordance for both technicians' spirograms. However, only one-third of those containing Certified Respiratory Technicians and Pulmonary Specialists exhibited this trend. In the majority of remaining pairs, greater levels of agreement corresponded to spirograms submitted by the highly trained technician. Only two (out of seven), both of which involved comparisons with Certified, Respiratory Technicians, displayed the reverse findings. Because results were discrepant, a positive relationship between strength of concordance and technician expertise appeared to exist on a pair-specific (as opposed to an expertise-specific) level. Comparable conclusions were derived from complementary spirograms (refer to Tables 4.9 and 4.10). In half of pairs representing each category combination, agreement did not differ across spirogram sets. Among the remaining half, levels strengthened with technician expertise. Consequently, a relationship between variations in technician effects and raters' levels of respiratory expertise was not evident.

ii) The Secondary Data Set

a) Comparisons Between Non-certified, Minimally Trained, Respiratory Research Assistants and Raters With Professional Respiratory Expertise

Similar to the Primary Data Set, no indication of a definite relationship between variances in observed "technician effects" and raters' respiratory expertise level was detected (refer to Tables 4.11 and 4.12). Best test spirograms submitted by both technicians produced analogous, broad ranges of kappa values in pairs that combined Non-certified, Minimally Trained, Respiratory Research Assistants with either Certified, Respiratory Technicians or Respiratory Epidemiologists. Changes in magnitude between stratified scores by technician expertise for common sets of raters were inconsistent across category comparisons. Concordance levels in seven out of nine pairs (77.8%) comprised of Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians did not significantly differ with technician expertise. However, among Non-certified, Minimally Trained, Respiratory Research Assistants and Respiratory Epidemiologists, only one-sixth of pairs displayed this trend. Moreover, half exhibited greater agreement for the highly trained technician's spirograms. In the remaining one-third, spirograms from the minimally trained technician generated stronger concordance. Among pairs who represented the two extremes in expertise categories (i.e., Non-certified, Minimally Trained, Respiratory Research Assistants and Pulmonary Specialists), all but one displayed equal degrees of agreement for each technician's spirograms. Therefore, beyond a pair-specific level, no consistent effect of technician expertise on agreement between Non-certified, Minimally Trained, Respiratory Research Assistants and raters with professional respiratory expertise was established. Similar conclusions were derived from complementary spirograms (refer to Tables 4.13 and 4.14). Among the majority of pairs, regardless of expertise, differences in agreement for spirograms from the two technicians were not statistically significant. With one exception (involving Non-certified, Minimally Trained, Respiratory Research Assistants and Pulmonary Specialists), no pair produced greater concordance for the highly trained technician's spirograms. Consequently, even across pairs who exhibited variances in concordance, a systematic "technician effect" was not found.

b) Comparisons Between Raters With Professional Respiratory Expertise

With respect to participants' best test spirograms, technician expertise did not appear to significantly influence agreement among raters whose professional respiratory expertise differed (refer to Tables 4.11 and 4.12). Results of at least two-thirds of pairs defining each expertise-based comparison revealed no difference in concordance between technicians' spirograms. However, across remaining pairs, tracings from the minimally trained technician consistently generated higher levels. These trends coincided with those established for complementary spirograms (refer to Tables 4.13 and 4.14). In two out of sixteen pairs, stronger concordance was associated with the minimally trained technician's tracings. Otherwise, no statistically significant variations in the degree of inter-rater reliability were observed.

iii) Summary Analysis of the Effect of Technician Expertise on Agreement Between Raters With Different Respiratory Expertise

In the Primary Data Set, regardless of spirogram type, mean "stratified" kappa values varied by no more than 0.18 indicating that apparent effects of technician expertise on average between-category concordance were not statistically significant[1]. Nevertheless, differences in summary scores generated by combinations of Non-certified, Minimally Trained, Respiratory Research Assistants and both Respiratory Epidemiologists and Pulmonary Specialists were only marginally non-significant (0.18 and 0.17, respectively) With respect to the Secondary Data Set, a maximum (marginally non-significant)

[1] A cut-off value of 0.20 was used in order to determine "statistical significance".

difference of 0.19 corresponded to category comparisons involving Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians. These three combinations, all involving raters with minimal and professional respiratory expertise, generated larger summary values for spirograms produced by the highly trained technician. Therefore, despite failure to meet statistical significance criteria (which should be interpreted with caution) their results were considered "clinically important".

4.3 Conclusions

A comparative analysis of kappa coefficients assessed variations in agreement regarding the acceptability of spirometric results associated with the respiratory expertise levels of raters and the training and experience of technicians.

Regardless of expertise category and spirogram type, agreement exhibited among raters whose respiratory expertise was similar did not exceed "moderate" levels. In addition to a lack of consistency in results across pairs within categories, no distinct correlation between strength of concordance and raters' level of respiratory expertise was detected. Findings are graphically illustrated in Figures 4.1 and 4.2. Also, observed effects of technician expertise on within category agreement were restricted both to the Primary Data Set and to categories representing the extremes in expertise. Between Pulmonary Specialists, stronger concordance coincided with best test spirograms from the highly trained technicians and with complementary spirograms from the minimally trained technician. Between Non-certified, Minimally Trained, Respiratory Research Assistants complementary spirograms generated the opposite pattern. This pattern, although not as distinct, was noted again in the Secondary Data Set. Therefore, effects of technician expertise on within-category agreement were deemed specific to spirogram type and rater expertise (refer to Figures 4.3 and 4.4).

The magnitude of agreement between raters with different levels of respiratory expertise was also pair-specific. Consequently, no definitive statement could be established regarding the effect of raters' respiratory expertise on inter-rater reliability. However, based on general patterns observed in kappa scores, concordance appeared to strengthen slightly as the difference between raters' levels of expertise decreased (refer to Figures 4.5 and 4.6). With respect to technician effects, no uniform trends across pairs representing common comparisons were recognized (refer to Figures 4.7 and 4.8). Thus, results did not conclusively indicate a correlation between technician expertise and agreement among raters with various levels of respiratory expertise.

In summary, analysis of kappa values revealed no systematic differences in concordance directly attributable to either rater or technician expertise.

Table 4.1 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Best Test Spirograms from First Test Sessions in the Primary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologists		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ •(95% C.I.)	κ (95% C.I.)	κ (95% C.I.	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I)	κ (95% C.I.)
Rater 1	n/a	0.21 (0.01 - 0.40)	0.36 (0.17 - 0.55)	0.14 (-0.06 - 0.33)	0.47 (0.29 - 0.64)	0.01 (-0.19 - 0.20)	0.47 (0.29 - 0.64)	0.58 (0.41 - 0.74)	0.39 (0.20 - 0.57)	0.20 (0.01 - 0.39)
Rater 2	-----	n/a	0.52 (0.35 - 0.70)	0.37 (0.19 - 0.56)	0.59 (0.43 - 0.75)	0.21 (0.01 - 0.41)	0.22 (0.02 - 0.41)	0.44 (0.26 - 0.62)	0.25 (0.06 - 0.45)	0.20 (0.01 - 0.39)
Rater 3	-----	-----	n/a	0.19 (-0.01 - 0.38)	0.44 (0.26 - 0.61)	0.11 (-0.09 - 0.31)	0.33 (0.14 - 0.52)	0.38 (0.19 - 0.57)	0.32 (0.12 - 0.51)	0.10 (-0.09 -
Rater 4	-----	-----	-----	n/a	0.51 (0.34 - 0.68)	0.29 (0.08 - 0.46)	0.33 (0.14 - 0.52)	0.26 (0.07 - 0.45)	0.36 (0.17 - 0.54)	0.31 (0.12 - 0.49)
Rater 5	-----	-----	-----	-----	n/a	0.41 (0.22 - 0.58)	0.51 (0.34 - 0.68)	0.63 (0.48 - 0.78)	0.61 (0.45 - 0.77)	0.53 (0.36 -0. 69)
Rater 6	-----	-----	-----	-----	-----	n/a	0.32 (0.13 - 0.51)	0.27 (0.07 - 0.47)	0.48 (0.30 - 0.66)	0.43 (0.25 - 0.61)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.58 (0.42 - 0.75)	0.52 (0.35 - 0.69)	0.37 (0.19 - 0.55)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.63 (0.47 - 0.79)	0.44 (0.26 - 0.62)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.50 (0.33 - 0.68)
Rater 10										n/a

•95% Confidence Interval

Table 4.2 Level of Agreement Between Individual Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Complementary Spirograms from First Test Sessions in the Primary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistant			Certified Respiratory Technician			Respiratory Epidemiologist		Pulmonary Specialist	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
Rater 1	n/a	0.18 (-0.01 - 0.38)	0.20 (0.02 - 0.40)	0.04 (-0.16 - 0.23)	0.37 (0.18 - 0.55)	0.20 (0.01 - 0.40)	0.42 (0.24 - 0.59)	0.41 (0.23 - 0.59)	0.35 (0.16 - 0.53)	0.18 (-0.01 - 0.38)
Rater 2	-----	n/a	0.45 (0.28 - 0.63)	0.46 (0.29 - 0.63)	0.49 (0.32 - 0.66)	0.22 (0.03 - 0.42)	0.32 (0.14 - 0.51)	0.33 (0.15 - 0.52)	0.27 (0.08 - 0.46)	0.12 (-0.08 - 0.46)
Rater 3	-----	-----	n/a	0.21 (0.01 - 0.40)	0.33 (0.14 - 0.52)	0.16 (-0.04 - 0.35)	0.20 (0.00 - 0.39)	0.22 (0.02 - 0.41)	0.12 (-0.08 - 0.31)	0.05 (-0.15 - 0.25)
Rater 4	-----	-----	-----	n/a	0.50 (0.32 - 0.67)	0.35 (0.16 - 0.54)	0.34 (0.15 - 0.52)	0.25 (0.05 - 0.44)	0.42 (0.25 - 0.60)	0.46 (0.29 - 0.63)
Rater 5	-----	-----	-----	-----	n/a	0.41 (0.23 - 0.59)	0.59 (0.43 - 0.75)	0.57 (0.41 - 0.73)	0.57 (0.41 - 0.73)	0.27 (0.08 - 0.46)
Rater 6	-----	-----	-----	-----	-----	n/a	0.47 (0.29 - 0.65)	0.58 (0.41 - 0.74)	0.39 (0.20 - 0.57)	0.55 (0.38 - 0.71)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.45 (0.27 - 0.62)	0.44 (0.27 - 0.62)	0.61 (0.45 - 0.77)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.50 (0.32 -0. 67)	0.60 (0.44 - 0.76)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	r/a	0.50 (0.32-0.67)
Rater 10										n/a

•95% Confidence Interval

Table 4.3 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Best Test Spirograms in the Secondary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologists		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ *(95% C.I.)	κ (95% C.I.)	κ (95% C.I)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)
Rater 1	n/a	0.45 (0.32 - 0.58)	0.52 (0.39 - 0.64)	0.03 (-0.11 - 0.17)	0.25 (0.12 - 0.39)	0.14 (-0.01 - 0.28)	0.31 (0.17 - 0.45)	0.26 (0.10 - 0.41)	0.29 (0.16 - 0.43)	0.24 (0.10 - 0.38)
Rater 2	-----	n/a	0.70 (0.60 - 0.81)	0.15 (0.01 - 0.29)	0.54 (0.42 - 0.66)	0.45 (0.32 - 0.58)	0.46 (0.33 - 0.59)	0.38 (0.24 - 0.53)	0.36 (0.23 - 0.50)	0.40 (0.27 - 0.53)
Rater 3	-----	-----	n/a	0.09 (-0.05 - 0.23)	0.41 (0.28 - 0.53)	0.29 (0.16 - 0.43)	0.38 (0.25 - 0.52)	0.36 (0.20 - 0.51)	0.31 (0.17 - 0.45)	0.31 (0.17 - 0.45)
Rater 4	-----	-----	-----	n/a	0.29 (0.15 - 0.43)	0.21 (0.07 - 0.34)	0.15 (0.01 - 0.29)	0.05 (-0.09 - 0.19)	0.16 (0.03 - 0.30)	0.18 (0.04 - 0.32)
Rater 5	-----	-----	-----	-----	n/a	0.51 (0.39 - 0.63)	0.52 (0.39 - 0.64)	0.30 (0.16 - 0.45)	0.51 (0.38 - 0.63)	0.47 (0.34 - 0.59)
Rater 6	-----	-----	-----	-----	-----	n/a	0.41 (0.27 - 0.54)	0.38 (0.23 - 0.52)	0.47 (0.34 - 0.60)	0.50 (0.38 - 0.63)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.48 (0.33 - 0.62)	0.66 (0.55 - 0.77)	0.33 (0.20 - 0.47)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.48 (0.33 - 0.62)	0.32 (0.17 - 0.46)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.40 (0.26 - 0.53)
Rater 10										n/a

*95% Confidence Interval

Table 4.4 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Complementary Spirograms in the Secondary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologists		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
	*(95% C.I.)	(95% C.I.)	(95% C.I)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)
Rater 1	n/a	0.33 (0.19 - 0.46)	0.43 (0.30 - 0.57)	0.02 (-0.12 - 0.16)	0.13 (-0.01 - 0.27)	0.04 (-0.10 - 0.18)	0.31 (0.17 - 0.46)	0.11 (-0.04 - 0.27)	0.10 (-0.04 - 0.24)	0.14 (0.00 - 0.28)
Rater 2	-----	n/a	0.61 (0.49 - 0.73)	0.18 (0.05 - 0.32)	0.43 (0.30 - 0.55)	0.26 (0.12 - 0.39)	0.45 (0.32 - 0.58)	0.24 (0.09 - 0.39)	0.32 (0.19 - 0.46)	0.33 (0.20 - 0.47)
Rater 3	-----	-----	n/a	0.11 (-0.03 - 0.25)	0.17 (0.03 - 0.30)	0.06 (-0.08 - 0.20)	0.48 (0.34 - 0.61)	0.28 (0.12 - 0.43)	0.29 (0.15 - 0.43)	0.29 (0.15 - 0.43)
Rater 4	-----	-----	-----	n/a	0.31 (0.18 - 0.45)	0.23 (0.10 - 0.37)	0.13 (-0.01 - 0.27)	0.04 (-0.10 - 0.17))	0.12 (-0.01 - 0.26)	0.14 (0.01 - 0.28)
Rater 5	-----	-----	-----	-----	n/a	0.48 (0.36 - 0.61)	0.39 (0.26 - 0.52)	0.18 (0.03 - 0.32)	0.52 (0.40 - 0.64)	0.40 (0.27 - 0.53)
Rater 6	-----	-----	-----	-----	-----	n/a	0.16 (0.01 - 0.30)	0.13 (-0.02 - 0.28)	0.34 (0.20 - 0.48)	0.34 (0.20 - 0.47)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.28 (0.12 - 0.44)	0.56 (0.44 - 0.69)	0.44 (0.31 - 0.58)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.27 (0.11 - 0.43)	0.14 (-0.01 - 0.30)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.45 (0.31 - 0.58)
Rater 10										n/a

*95% Confidence Interval

Table 4.5 Average Kappa Values Within and Between Groups of Raters (i.e., Expertise Levels) From First Spirometry Sessions of the Primary Data Set

	Non-Certified, Minimally Trained, Respiratory Research Assistants (RA)		Certified Respiratory Technicians (CRT)		Respiratory Epidemiologists (RE)		Pulmonary Specialists (PS)	
	Best Test Spirogram	Complementary Spirogram	Best Test Spirogram	Complementary Spirogram	Best Test Spirogram	Complementary Spirogram	Best Test Spirogram	Complementary Spirogram
	κ	κ	κ	κ	κ	κ	κ	κ
RA	0.36	0.28	0.28	0.28	0.40	0.32	0.24	0.18
CRT	-----	-----	0.41	0.42	0.26	0.47	0.30	0.44
RE	-----	-----	-----	-----	0.58	0.45	0.49	0.54
PS	-----	-----	-----	-----	-----	-----	0.50	0.50

Table 4.6 Average Kappa Values Within and Between Groups of Raters (i.e., Expertise Levels) in the Secondary Data Set

	Non-Certified, Minimally Trained, Respiratory Research Assistants (RA)		Certified Respiratory Technicians (CRT)		Respiratory Epidemiologists (RE)		Pulmonary Specialists (PS)	
	Best Test Spirogram	Complementary Spirogram	Best Test Spirogram	Complementary Spirogram	Best Test Spirogram	Complementary Spirogram	Best Test Spirogram	Complementary Spirogram
	κ	κ	κ	κ	κ	κ	κ	κ
RA	0.56	0.46	0.26	0.16	0.36	0.31	0.32	0.25
CRT	-----	-----	0.34	0.34	0.30	0.17	0.38	0.31
RE	-----	-----	-----	-----	0.48	0.28	0.45	0.35
PS	-----	-----	-----	-----	-----	-----	0.40	0.45

Table 4.7 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Best Test Spirograms from First Test Sessions Administered By the Minimally Trained and Experienced Technician in the Primary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologist		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ *(95% C.I.)	κ (95% C.I.)	κ (95% C.I)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)
Rater 1	n/a	0.31 (0.01 - 0.60)	0.28 (-0.01 - 0.57)	0.27 (-0.03 - 0.58)	0.41 (0.11 - 0.71)	0.02 (-0.27 - 0.30)	0.43 (0.15 - 0.72)	0.56 (0.30 - 0.83)	0.34 (0.05 - 0.63)	0.09 (-0.21 - 0.39)
Rater 2	-----	n/a	0.43 (0.16 - 0.71)	0.53 (0.25 - 0.80)	0.66 (0.42 - 0.91)	0.16 (-0.13 - 0.46)	0.06 (-0.23 - 0.36)	0.46 (0.18 - 0.74)	0.23 (-0.07 - 0.53)	0.08 (-0.22 - 0.38)
Rater 3	-----	-----	n/a	0.34 (0.05 - 0.63)	0.34 (0.05 - 0.63)	0.11 (-0.18 - 0.40)	0.17 (-0.12 - 0.47)	0.30 (0.01 - 0.59)	0.33 (0.05 - 0.62)	-0.06 (-0.34 - 0.22)
Rater 4	-----	-----	-----	n/a	0.62 (0.35 - 0.89)	0.34 (0.05 - 0.63)	0.31 (0.01 - 0.61)	0.44 (0.15 - 0.72)	0.48 (0.20 - 0.76)	0.23 (-0.08 - 0.53)
Rater 5	-----	-----	-----	-----	n/a	0.34 (0.05 - 0.63)	0.37 (0.07 - 0.67)	0.67 (0.42 - 0.91)	0.63 (0.37 - 0.88)	0.28 (-0.03 - 0.59)
Rater 6	-----	-----	-----	-----	-----	n/a	0.38 (0.10 - 0.66)	0.22 (-0.07 - 0.51)	0.45 (0.18 - 0.73)	0.23 (-0.07 - 0.52)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.59 (0.34 - 0.84)	0.54 (0.28 - 0.80)	0.21 (-0.09 - 0.51)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.67 (0.44 - 0.90)	0.42 (0.14 - 0.71)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.38 (0.10 - 0.65)
Rater 10										n/a

*95% Confidence Interval

Table 4.8 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Best Test Spirograms from First Test Sessions Administered By the Highly Trained and Experienced Technician in the Primary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologist		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ *(95% C.I.)	κ (95% C.I.)	κ (95% C.I)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)	κ (95% C.I.)
Rater 1	n/a	0.12 (-0.15 - 0.39)	0.45 (0.20 - 0.69)	0.04 (-0.23 - 0.31)	0.52 (0.29 - 0.75)	0.00 (-0.27 - 0.27)	0.50 (0.27 - 0.74)	0.59 (0.36 - 0.82)	0 44 (0.19 - 0.69)	0.31 (0.05 - 0.56)
Rater 2	-----	n/a	0.61 (0.39 - 0.83)	0.25 (-0.01 - 0.51)	0.52 (0.29 - 0.75)	0.25 (-0.02 - 0.52)	0.35 (0.09 - 0.60)	0.43 (0.17 - 0.68)	0.28 (0.17 - 0.68)	0.31 (0.05 - 0.56)
Rater 3	-----	-----	n/a	0.02 (-0.25 - 0.29)	0.52 (0.29 - 0.75)	0.11 (-0.16 - 0.38)	0.43 (0.18 - 0.67)	0.44 (0.18 - 0.67)	0.29 (0.03 - 0.55)	0.23 (-0.03 - 0.50)
Rater 4	-----	-----	-----	n/a	0.38 (0.13 - 0.630	0.21 (-0.05 - 0.47)	0.33 (0.07 - 0.59)	0.13 (-0.14 - 0.39)	0.25 (-0.01 - 0.51)	0.37 (0.12 - 0.62)
Rater 5	-----	-----	-----	-----	n/a	0.45 (0.21 - 0.70)	0.63 (0.41 - 0.84)	0.60 (0.38 - 0.81)	0.59 (0.38 - 0.81)	0.74 (0.56 - 0.92)
Rater 6	-----	-----	-----	-----	-----	n/a	0.27 (0.00 - 0.53)	0.31 (0.04 - 0.58)	0.50 (0.27 - 0.74)	0.61 (0.39 - 0.83)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.58 (0.35 - 0.80)	0.50 (0.27 - 0.74)	0.51 (0.27 - 0.74)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.59 (0.36 - 0.82)	0.46 (0.21 - 0.70)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.61 (0.40 - 0.83)
Rater 10										n/a

*95% Confidence Interval

Table 4.9 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Complementary Spirograms from First Test Sessions Administered By the Minimally Trained and Experienced Technician in the Primary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologist		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
	*(95% C.I.)	(95% C.I.)	(95% C.I)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)
Rater 1	n/a	0.14 (-0.14 - 0.43)	0.06 (-0.22 - 0.35)	0.13 (-0.13 - 0.42)	0.30 (0.02 - 0.58)	0.22 (-0.06 - 0.50)	0.42 (0.16 - 0.69)	0.32 (0.05 - 0.60)	0.53 (0.29 - 0.78)	0.44 (0.18 - 0.70)
Rater 2	-----	n/a	0.16 (-0.13 - 0.44)	0.10 (-0.18 - 0.39)	0.28 (0.00 - 0.55)	0.07 (-0.22 - 0.35)	-0.06 (-0.35 - 0.22)	0.25 (-0.03 - 0.53)	0.19 (-0.09 - 0.47)	-0.04 (-0.32 - 0.24)
Rater 3	-----	-----	n/a	0.25 (-0.03 - 0.53)	0.19 (-0.09 - 0.47)	-0.04 (-0.32 - 0.24)	0.15 (-0.14 - 0.43)	0.10 (-0.19 - 0.39)	-0.04 (-0.33 - 0.24)	0.19 (-0.09 - 0.47)
Rater 4	-----	-----	-----	n/a	0.51 (0.26 - 0.75)	0.52 (0.26 - 0.77)	0.25 (-0.03 - 0.52)	0.40 (0.14 - 0.67)	0.53 (0.29 - 0.77)	0.40 (0.13 - 0.66)
Rater 5	-----	-----	-----	-----	n/a	0.42 (0.16 - 0.68)	0.46 (0.20 - 0.72)	0.58 (0.34 - 0.81)	0.63 (0.40 - 0.85)	0.42 (0.16 - 0.68)
Rater 6	-----	-----	-----	-----	-----	n/a	0.31 (0.04 - 0.58)	0.44 (0.18 - 0.70)	0.74 (0.54 - 0.94)	0.43 (0.16 - 0.69)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.58 (0.34 - 0.81)	0.37 (0.10 - 0.63)	0.33 (0.06 - 0.60)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.62 (0.39 - 0.84)	0.49 (0.24 - 0.74)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.70 (0.50 - 0.90)
Rater 10										n/a

*95% Confidence Interval

Table 4.10 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Complementary Spirograms from First Test Sessions Administered By the Highly Trained and Experienced Technician in the Primary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologist		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
	*(95% C.I.)	(95% C.I.)	(95% C.I)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)
Rater 1	n/a	0.20 (-0.07 - 0.46)	0.39 (0.14 - 0.64)	0.00 (-0.27 - 0.27)	0.40 (0.16 - 0.65)	0.17 (-0.10 - 0.44)	0.40 (0.15 - 0.65)	0.42 (0.17 - 0.67)	0.42 (0.18 - 0.67)	0.24 (-0.02 - 0.51)
Rater 2	-----	n/a	0.58 (0.36 - 0.80)	0.41 (0.16 - 0.65)	0.52 (0.29 - 0.75)	0.28 (0.01 - 0.55)	0.51 (0.28 - 0.74)	0.37 (0.12 - 0.63)	0.46 (0.21 - 0.70)	0.28 (0.02 - 0.54)
Rater 3	-----	-----	n/a	0.15 (-0.12 - 0.41)	0.48 (0.24 - 0.71)	0.33 (0.07 - 0.59)	0.25 (-0.01 - 0.51)	0.34 (0.08 - 0.60)	0.27 (0.01 - 0.53)	0.24 (-0.02 - 0.51)
Rater 4	-----	-----	-----	n/a	0.50 (0.26 - 0.73)	0.23 (-0.04 - 0.49)	0.43 (0.18 - 0.67)	0.14 (-0.13 - 0.40)	0.34 (0.09 - 0.60)	0.51 (0.28 - 0.74)
Rater 5	-----	-----	-----	-----	n/a	0.40 (0.14 - 0.64)	0.70 (0.50 - 0.89)	0.60 (0.38 - 0.82)	0.52 (0.29 - 0.75)	0.70 (0.29 - 0.75)
Rater 6	-----	-----	-----	-----	-----	n/a	0.23 (-0.04 - 0.49)	0.48 (0.23 - 0.74)	0.42 (0.16 - 0.68)	0.35 (0.10 - 0.61)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.52 (0.28 - 0.75)	0.52 (0.28 - 0.75)	0.55 (0.32 - 0.77)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.60 (0.38 - 0.82)	0.51 (0.27 - 0.74)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.51 (0.27 - 0.74)
Rater 10										n/a

*95% Confidence Interval

Table 4.11 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Each Participants' Best Test Spirograms from Test Sessions Administered By Minimally Trained and Experienced Technicians in the Secondary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologists		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
	*(95% C.I.)	(95% C.I.)	(95% C.I)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)
Rater 1	n/a	0.41 (0.22 - 0.59)	0.39 (0.21 - 0.58)	-0.02 (-0.21 - 0.18)	0.27 (0.09 - 0.46)	0.17 (-0.02 - 0.37)	0.21 (0.02 - 0.41)	0.29 (0.09 - 0.50)	0.26 (0.07 - 0.45)	0.19 (-0.01 - 0.38)
Rater 2	-----	n/a	0.70 (0.55 - 0.85)	0.10 (-0.10 - 0.30)	0.54 (0.37 - 0.70)	0.54 (0.37 - 0.70)	0.38 (0.20 - 0.56)	0.50 (0.31 - 0.68)	0.34 (0.15 - 0.53)	0.52 (0.35 - 0.69)
Rater 3	-----	-----	n/a	0.04 (-0.16 - 0.24)	0.43 (0.25 - 0.61)	0.41 (0.23 - 0.59)	0.29 (0.09 - 0.46)	0.47 (0.27 - 0.66)	0.26 (0.07 - 0.45)	0.35 (0.16 - 0.54)
Rater 4	-----	-----	-----	n/a	0.27 (0.06 - 0.47)	0.20 (0.00 - 0.40)	0.16 (-0.04 - 0.36)	0.03 (-0.16 - 0.23)	0.18 (-0.02 - 0.38)	0.16 (-0.04 - 0.35)
Rater 5	-----	-----	-----	-----	n/a	0.58 (0.42- 0.73)	0.53 (0.37- 0.70)	0.31 (0.12 - 0.50)	0.54 (0.37 - 0.71)	0.48 (0.30 - 0.65)
Rater 6	-----	-----	-----	-----	-----	n/a	0.50 (0.33 - 0.67)	0.45 (0.26 - 0.63)	0.44 (0.10 - 0.48)	0.72 (0.58 - 0.86)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.43 (0.25 - 0.62)	0.62 (0.47 - 0.77)	0.37 (0.18 - 0.55)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.42 (0.23 - 0.61)	0.41 (0.21 - 0.60)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.40 (0.22 - 0.58)
Rater 10										n/a

*95% Confidence Interval

Table 4.12 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Best Test Spirograms from Test Sessions Administered By the Highly Trained and Experienced Technician in the Secondary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologists		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
	*(95% C.I.)	(95% C.I.)	(95% C.I)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)
Rater 1	n/a	0.50 (0.31 - 0.68)	0.65 (0.48 - 0.81)	0.07 (-0.13 - 0.27)	0.20 (0.00 - 0.40)	0.06 (-0.14 - 0.26)	0.41 (0.21 - 0.62)	0.19 (-0.04 - 0.42)	0.28 (0.06 - 0.50)	0.29 (0.09 - 0.49)
Rater 2	-----	n/a	0.70 (0.55 - 0.85)	0.19 (0.00 - 0.38)	0.53 (0.35 - 0.71)	0.33 (0.12 - 0.53)	0.54 (0.34 - 0.73)	0.23 (0.00 - 0.45)	0.36 (0.15 - 0.57)	0.25 (0.05 - 0.45)
Rater 3	-----	-----	n/a	0.13 (-0.07 - 0.32)	0.36 (0.16 - 0.55)	0.13 (-0.08 - 0.33)	0.48 (0.28 - 0.69)	0.20 (-0.03 - 0.43)	0.35 (0.13 - 0.57)	0.26 (0.05 - 0.46)
Rater 4	-----	-----	-----	n/a	0.27 (0.08 - 0.46)	0.18 (-0.02 - 0.37)	0.10 (-0.09 - 0.30)	0.04 (-0.15 - 0.24)	0.10 (-0.10 - 0.30)	0.18 (-0.01 - 0.38)
Rater 5	-----	-----	-----	-----	n/a	0.36 (0.16 - 0.56)	0.42 (0.21 - 0.62)	0.22 (0.00 - 0.44)	0.35 (0.13 - 0.56)	0.42 (0.23 - 0.61)
Rater 6	-----	-----	-----	-----	-----	n/a	0.16 (-0.05 - 0.38)	0.18 (-0.06 - 0.42)	0.37 (0.14 - 0.59)	0.19 (-0.02 - 0.40)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.48 (0.21 - 0.74)	0.61 (0.39 - 0.83)	0.23 (0.01 - 0.44)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.48 (0.19 - 0.78)	0.14 (-0.08 - 0.37)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.30 (0.08 - 0.53)
Rater 10										n/a

*95% Confidence Interval

Table 4.13 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Complementary Spirograms from Test Sessions Administered By Minimally Trained and Experienced Technicians in the Secondary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologists		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
	*(95% C.I.)	(95% C.I.)	(95% C.I)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)
Rater 1	n/a	0.25 (0.05 - 0.45)	0.31 (0.10 - 0.52)	0.06 (-0.14 - 0.25)	0.13 (-0.07 - 0.33)	0.10 (-0.10 - 0.30)	0.29 (0.08 - 0.50)	0.20 (-0.03 - 0.44)	0.05 (-0.14 - 0.25)	0.17 (-0.03 - 0.37)
Rater 2	-----	n/a	0.49 (0.31 - 0.67)	0.21 (0.02 - 0.41)	0.45 (0.27 - 0.62)	0.35 (0.16 - 0.54)	0.43 (0.24 - 0.61)	0.20 (-0.01 - 0.41)	0.31 (0.12 - 0.50)	0.47 (0.29 - 0.65)
Rater 3	-----	-----	n/a	0.13 (-0.07 - 0.32)	0.20 (0.01 - 0.40)	0.17 (-0.03 - 0.37)	0.44 (0.24 - 0.63)	0.26 (0.03 - 0.49)	0.30 (0.10 - 0.49)	0.33 (0.13 - 0.52)
Rater 4	-----	-----	-----	n/a	0.27 (0.07 - 0.47)	0.18 (-0.01 - 0.38)	0.10 (-0.09 - 0.30)	0.02 (-0.18 - 0.22)	0.08 (-0.12 - 0.27)	0.13 (-0.06 - 0.33)
Rater 5	-----	-----	-----	-----	n/a	0.47 (0.29 - 0.64)	0.38 (0.19 - 0.56)	0.14 (-0.06 - 0.34)	0.51 (0.34 - 0.68)	0.35 (0.16 - 0.53)
Rater 6	-----	-----	-----	-----	-----	n/a	0.28 (0.08 - 0.47)	0.14 (-0.07 - 0.35)	0.38 (0.19 - 0.56)	0.41 (0.23 - 0.59)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.31 (0.08 - 0.53)	0.54 (0.37 - 0.71)	0.40 (0.21 - 0.59)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.24 (0.03 - 0.45)	0.16 (-0.05 - 0.37)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.45 (0.28 - 0.63)
Rater 10										n/a

*95% Confidence Interval

Table 4.14 Level of Agreement Between Raters (Expressed As Cohen's Kappa) for the Acceptability of Participants' Complementary Spirograms From Test Sessions Administered By the Highly Trained and Experienced Technician in the Secondary Data Set

	Non-Certified, Minimally Trained Respiratory Research Assistants			Certified Respiratory Technicians			Respiratory Epidemiologists		Pulmonary Specialists	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
	*(95% C.I.	(95% C.I.)	(95% C.I)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)	(95% C.I.)
Rater 1	n/a	0.41 (0.21 - 0.58)	0.53 (0.35 - 0.71)	0.00 (-0.20 - 0.20)	0.16 (-0.04 - 0.35)	0.00 (-0.20 - 0.20)	0.34 (0.14 - 0.53)	0.05 (-0.16 - 0.25)	0.20 (-0.01 - 0.41)	0.13 (-0.07 - 0.33)
Rater 2	-----	n/a	0.72 (0.58 - 0.86)	0.15 (-0.04 - 0.35)	0.41 (0.22 - 0.59)	0.16 (-0.04 - 0.35)	0.47 (0.29 - 0.66)	0.29 (0.08 - 0.50)	0.34 (0.14 - 0.54)	0.19 (-0.01 - 0.39)
Rater 3	-----	-----	n/a	0.10 (-0.10 - 0.29)	0.15 (-0.05 - 0.35)	-0.04 (-0.23 - 0.16)	0.52 (0.34 - 0.70)	0.29 (0.09 - 0.48)	0.31 (0.11 - 0.52)	0.27 (0.07 - 0.47)
Rater 4	-----	-----	-----	n/a	0.32 (0.14 - 0.51)	0.26 (0.07 - 0.45)	0.16 (-0.03 - 0.36)	0.05 (-0.14 - 0.25)	0.14 (-0.06 - 0.33)	0.14 (-0.06 - 0.33)
Rater 5	-----	-----	-----	-----	n/a	0.49 (0.30 - 0.67)	0.41 (0.22 - 0.60)	0.23 (0.01 - 0.44)	0.49 (0.30 - 0.68)	0.43 (0.25 - 0.62)
Rater 6	-----	-----	-----	-----	-----	n/a	0.02 (-0.18 - 0.21)	0.11 (-0.10 - 0.32)	0.27 (0.06 - 0.48)	0.24 (0.04 - 0.44)
Rater 7	-----	-----	-----	-----	-----	-----	n/a	0.26 (0.03 - 0.49)	0.59 (0.40 - 0.78)	0.49 (0.30 - 0.67)
Rater 8	-----	-----	-----	-----	-----	-----	-----	n/a	0.33 (0.07 - 0.58)	0.13 (-0.09 - 0.34)
Rater 9	-----	-----	-----	-----	-----	------	-----	-----	n/a	0.41 (0.20 - 0.61)
Rater 10										n/a

*95% Confidence Interval

Table 4.15 Average Kappa Values Within and Between Groups of Raters (i.e., Expertise Levels) Stratified By Technician Expertise for Participants' First Spirometry Sessions in the Primary Data Set

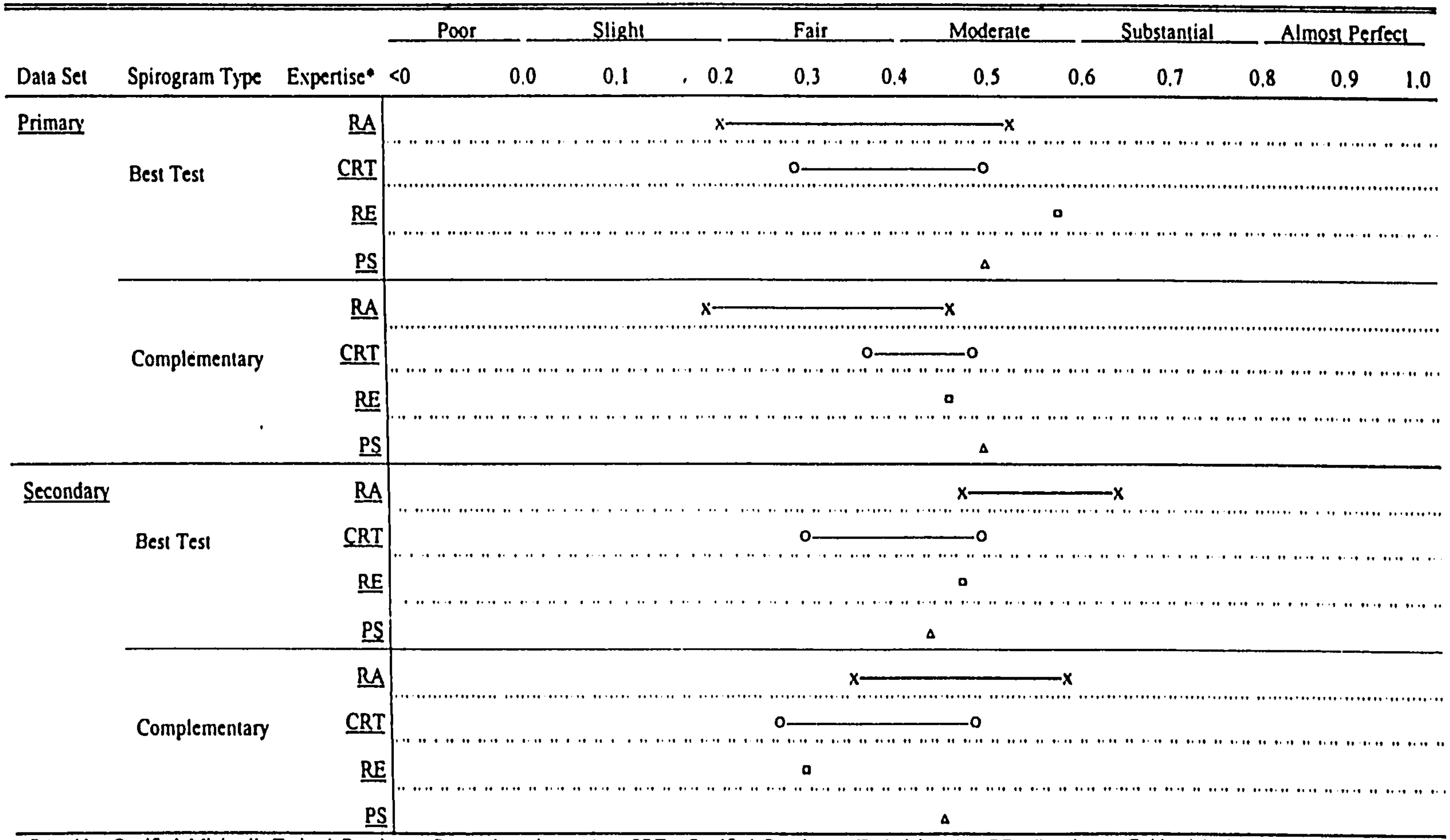
	Non-Certified, Minimally Trained, Respiratory Research Assistants (RA)				Certified Respiratory Technicians (CRT)				Respiratory Epidemiologists (RE)				Pulmonary Specialists (PS)			
	Minimally Trained Technician		Highly Trained Technician		Minimally Trained Technician		Highly Trained Technician		Minimally Trained Technician		Highly Trained Technician		Minimally Trained Technician		Highly Trained Technician	
	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
RA	0.34	0.12	0.39	0.39	0.32	0.17	0.25	0.30	0.33	0.20	0.46	0.38	0.17	0.15	0.31	0.32
CRT	-----	-----	-----	-----	0.43	0.48	0.35	0.38	0.40	0.41	0.38	0.43	0.38	0.53	0.51	0.43
RE	-----	-----	-----	-----	-----	-----	-----	-----	0.59	0.58	0.58	0.52	0.46	0.45	0.52	0.55
PS	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	0.38	0.70	0.61	0.51

* Complementary

Table 4.16 Average Kappa Values Within and Between Groups of Raters (i.e., Expertise Levels) Stratified By Technician Expertise in the Secondary Data Set

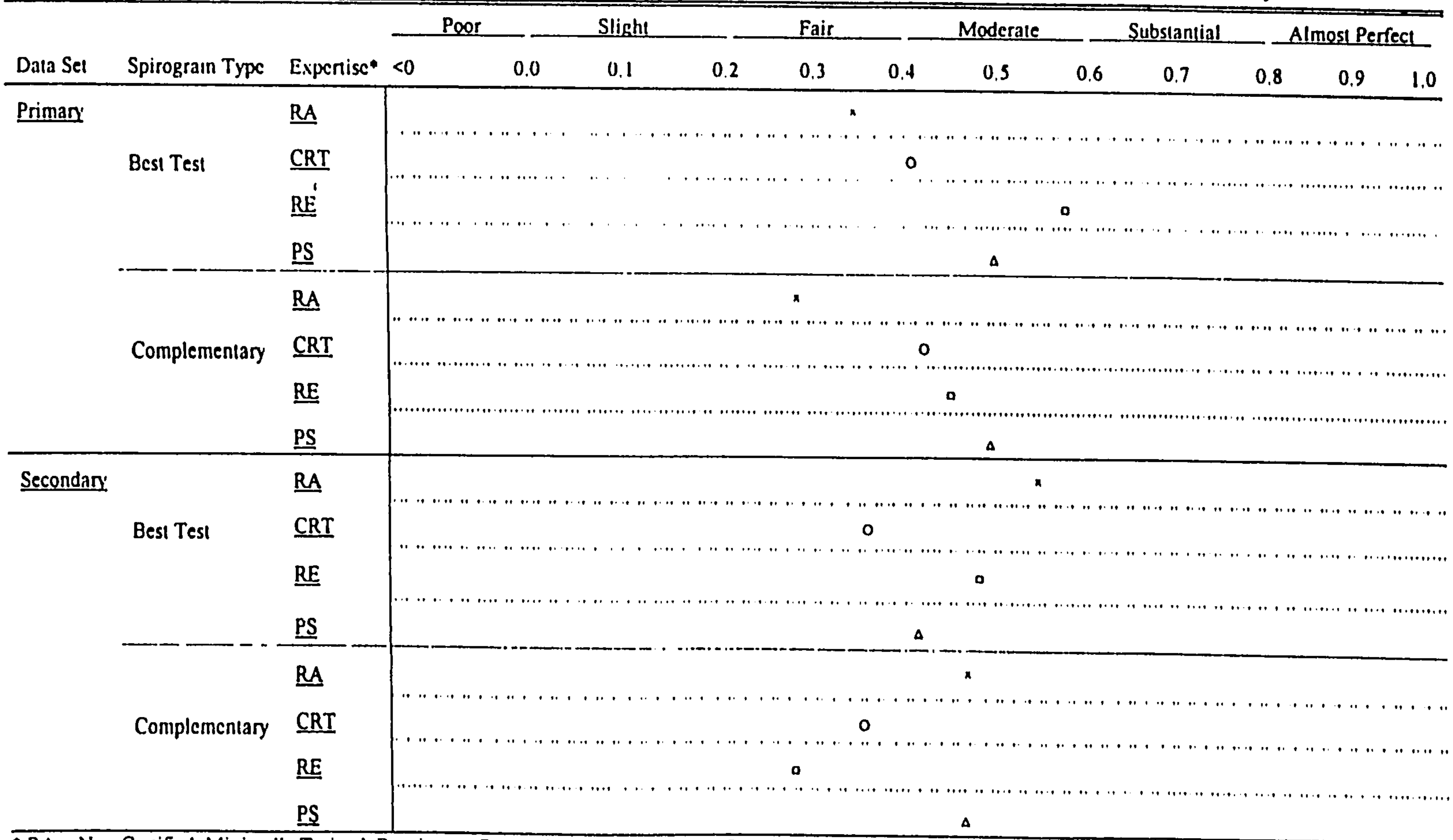
	Non-Certified, Minimally Trained, Respiratory Research Assistants (RA)				Certified Respiratory Technicians (CRT)				Respiratory Epidemiologists (RE)				Pulmonary Specialists (PS)			
	Minimally Trained Technician		Highly Trained Technician		Minimally Trained Technician		Highly Trained Technician		Minimally Trained Technician		Highly Trained Technician		Minimally Trained Technician		Highly Trained Technician	
	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirorgram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram	Best Test Spirogram	Comp* Spirogram
	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ	κ
RA	0.50	0.35	0.62	0.55	0.41	0.20	0.22	0.12	0.36	0.30	0.34	0.33	0.32	0.27	0.30	0.24
CRT	-----	-----	-----	-----	0.35	0.31	0.27	0.36	0.33	0.18	0.19	0.16	0.42	0.31	0.27	0.29
RE	-----	-----	-----	-----	-----	-----	-----	-----	0.43	0.31	0.48	0.26	0.46	0.34	0.37	0.39
PS	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	0.40	0.45	0.30	0.41

Figure 4.1 Range of Kappa Values Within Expertise Levels for Both Spirogram Types in the Primary and Secondary Data Sets



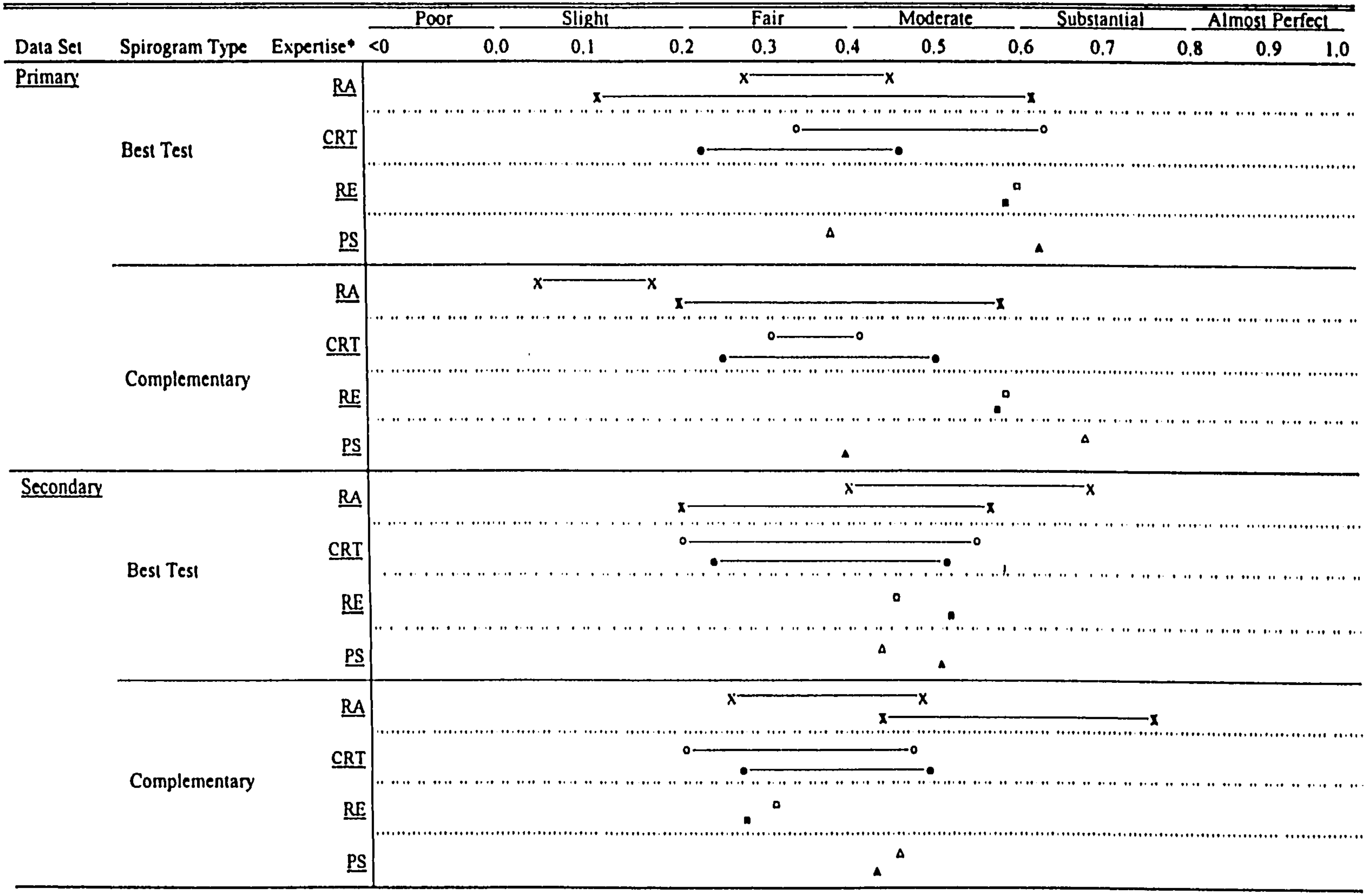
* RA = Non-Certified, Minimally Trained, Respiratory Research Assistants (x) ; CRT = Certified, Respiratory Technicians (o); RE = Respiratory Epidemiologists (▫);
PS = Pulmonary Specialists (▵)

Figure 4.2 Range of Average Kappa Values Within Expertise Levels for Both Spirogram Types in the Primary and Secondary Data Sets



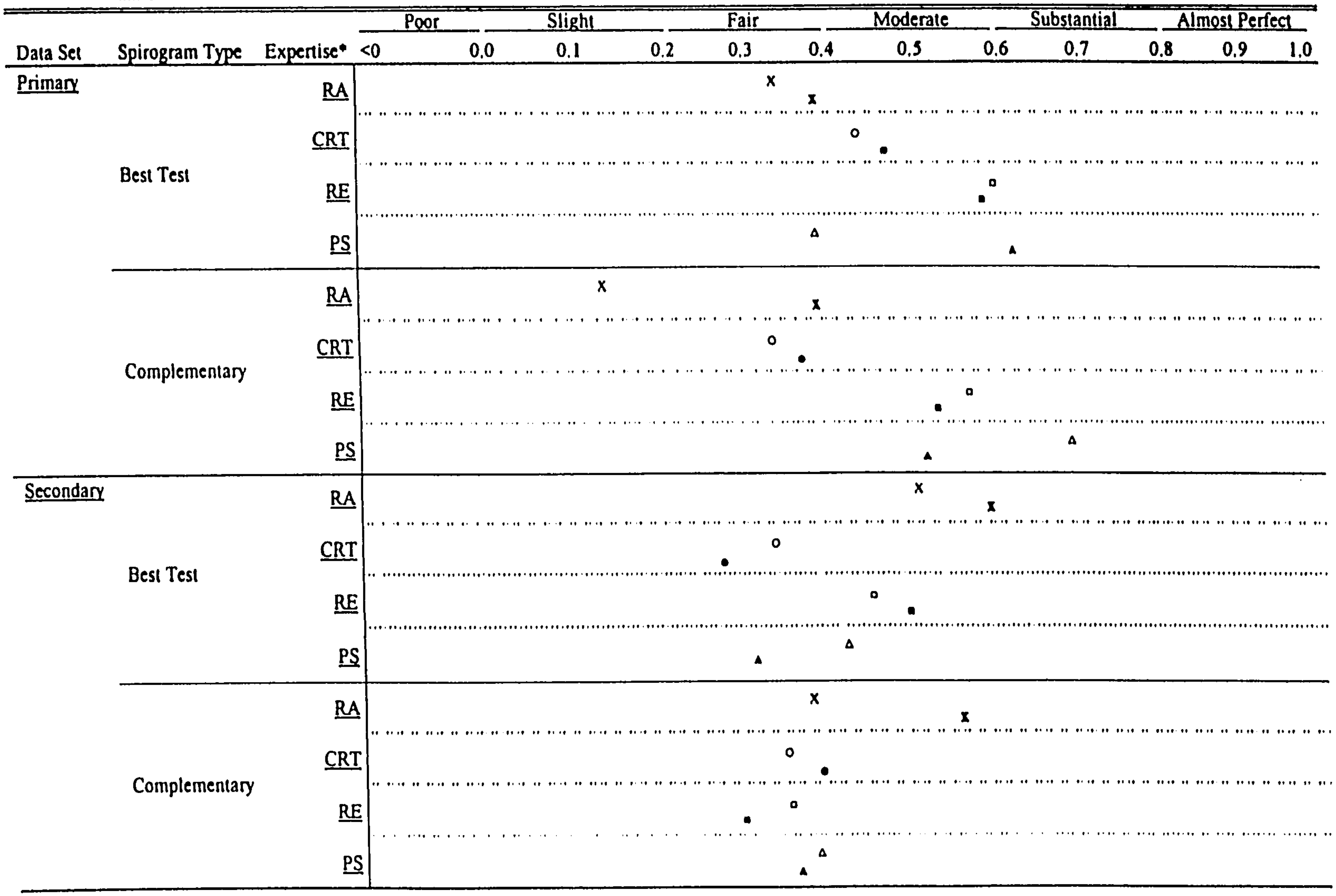
* RA = Non-Certified, Minimally Trained, Respiratory Research Assistants; CRT = Certified, Respiratory Technicians; RE = Respiratory Epidemiologists; PS = Pulmonary Specialists

Figure 4.3 Range of Kappa Values Stratified By Technician Expertise Within Expertise Levels for Both Spirogram Types in the Primary and Secondary Data Sets



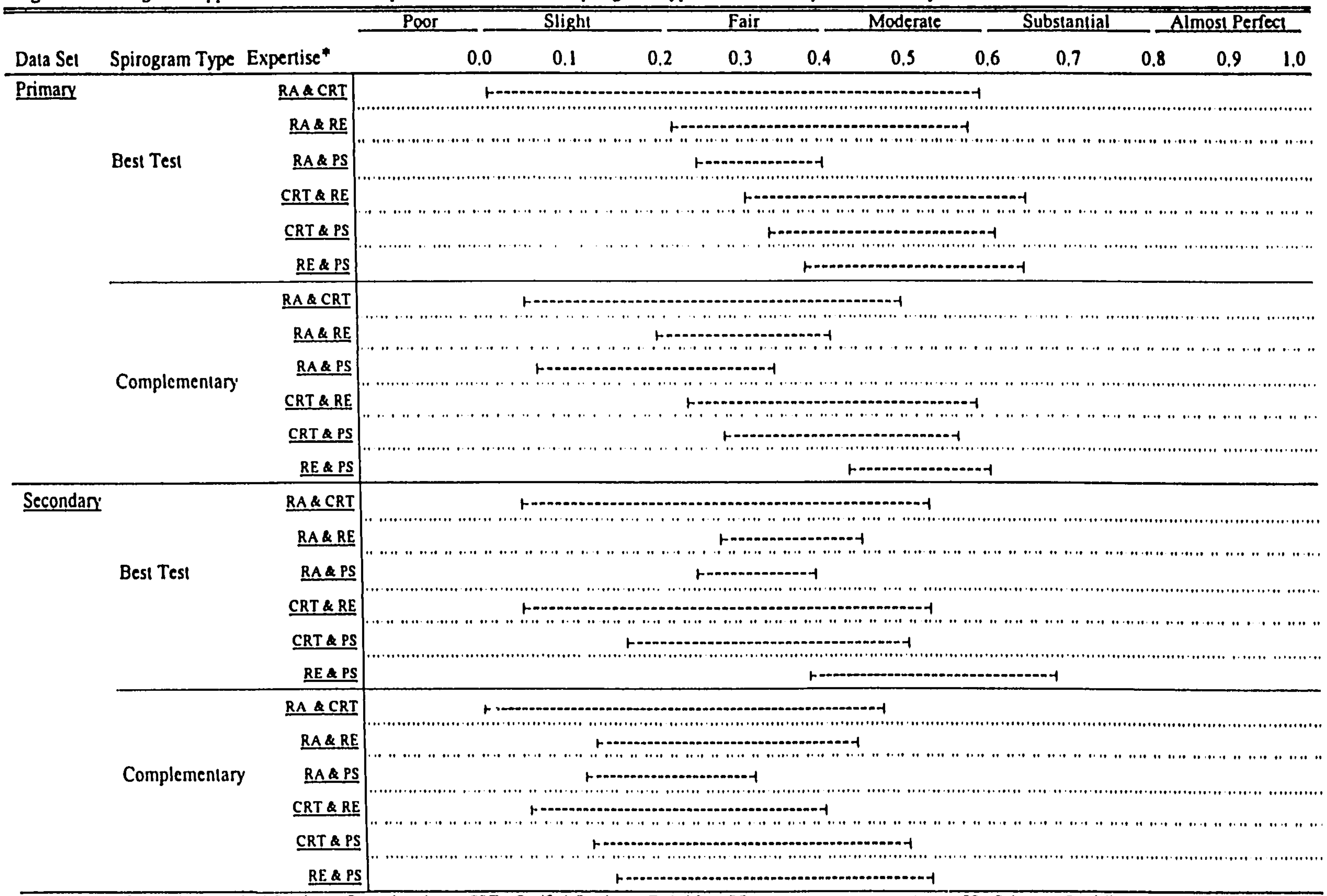
* RA = Non-Certified, Minimally Trained, Respiratory Research Assistants (x); CRT = Certified, Respiratory Technicians (o); RE = Respiratory Epidemiologists (▵) PS = Pulmonary Specialists (▵)
open shapes = based on spirograms from the minimally trained and experienced technician
solid or bold shapes = based on spirograms from the highly trained and experienced technician

Figure 4.4 Range of Average Kappa Values Stratified By Technician Expertise Within Expertise Levels for Both Spirogram Types in the Primary and Secondary Data Sets



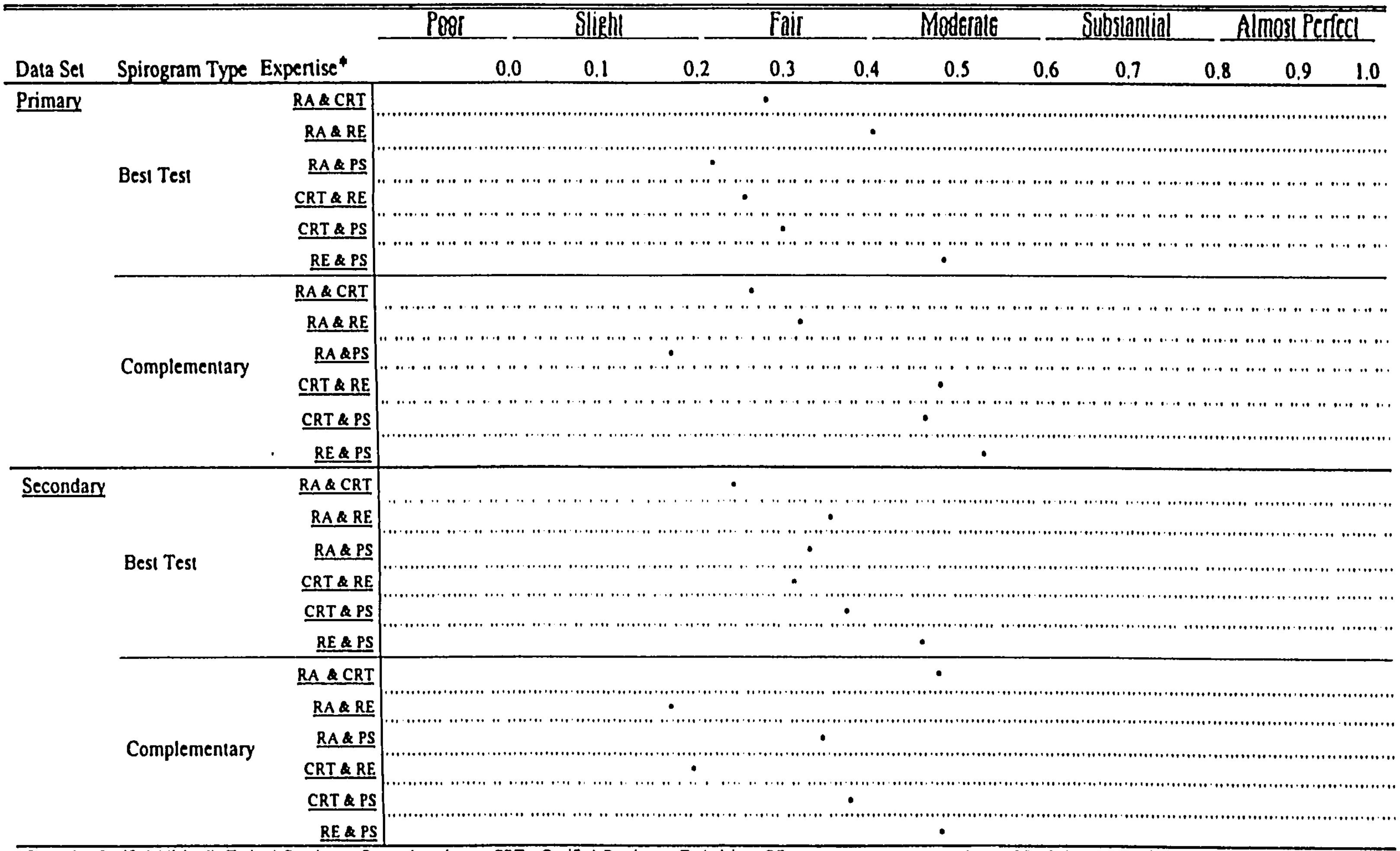
* RA = Non-Certified, Minimally Trained, Respiratory Research Assistants (x); CRT = Certified, Respiratory Technicians (o); RE = Respiratory Epidemiologists (□); PS = Pulmonary Specialists (△)
open shapes = Based on spirograms from the minimally trained and experienced technician
solid or bold shapes = Based on spirograms from the highly trained and experienced technician

Figure 4.5 Range of Kappa Values Between Expertise Levels for Both Spirogram Types in the Primary and Secondary Data Sets



* RA = Non-Certified, Minimally Trained, Respiratory Research Assistants; CRT = Certified, Respiratory Technicians; RE = Respiratory Epidemiologists ; PS = Pulmonary Specialists

Figure 4.6 Range of Average Kappa Values Between Expertise Levels for Both Spirogram Types in the Primary and Secondary Data Sets



* RA = Non-Certified, Minimally Trained, Respiratory Research Assistants; CRT = Certified, Respiratory Technicians; RE = Respiratory Epidemiologists ı; PS = Pulmonary Specialists

Figure 4.7 Range of Kappa Values Stratified By Technician Expertise Between Expertise Levels for Both Spirogram Types in the Primary and Secondary Data Sets

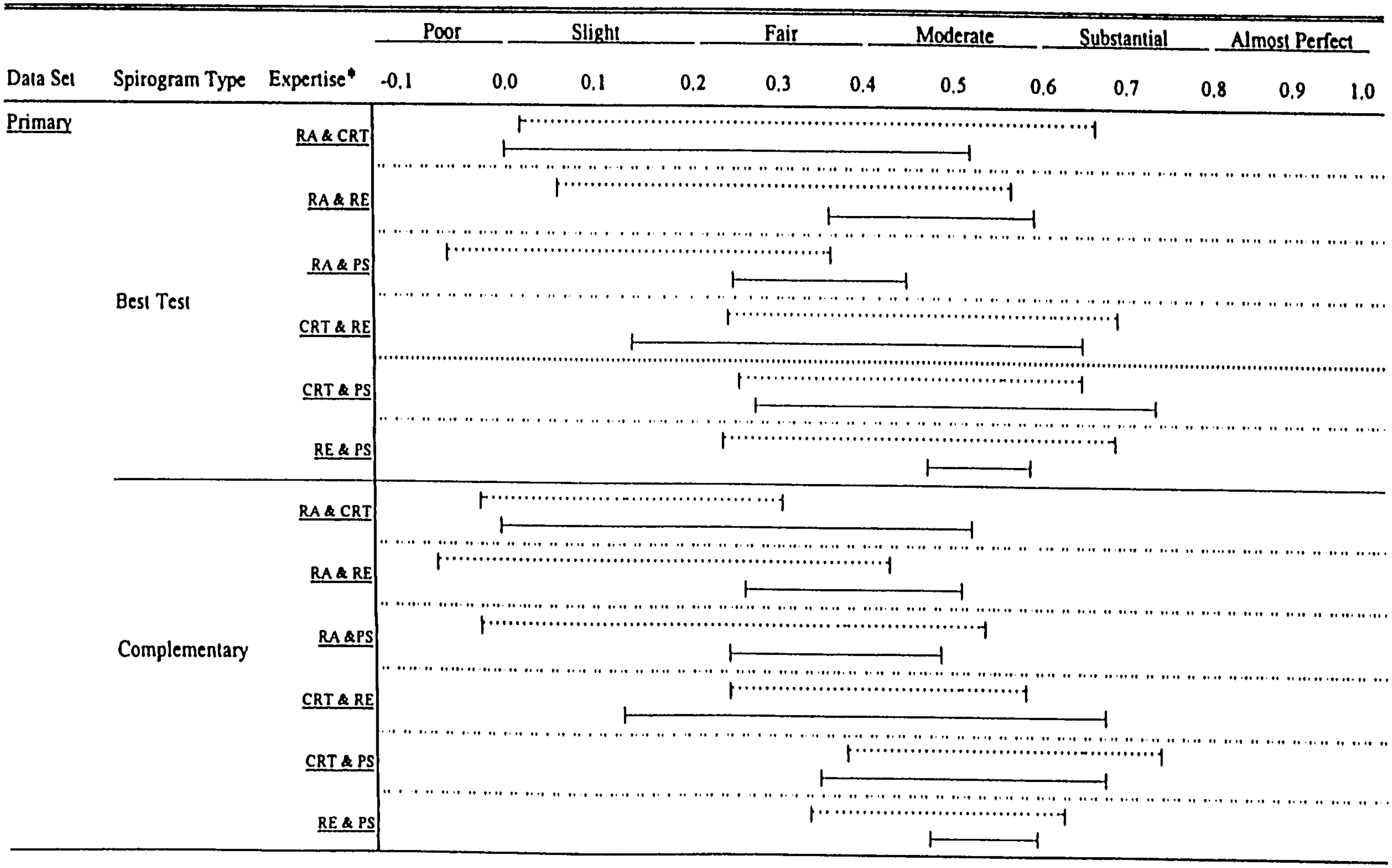
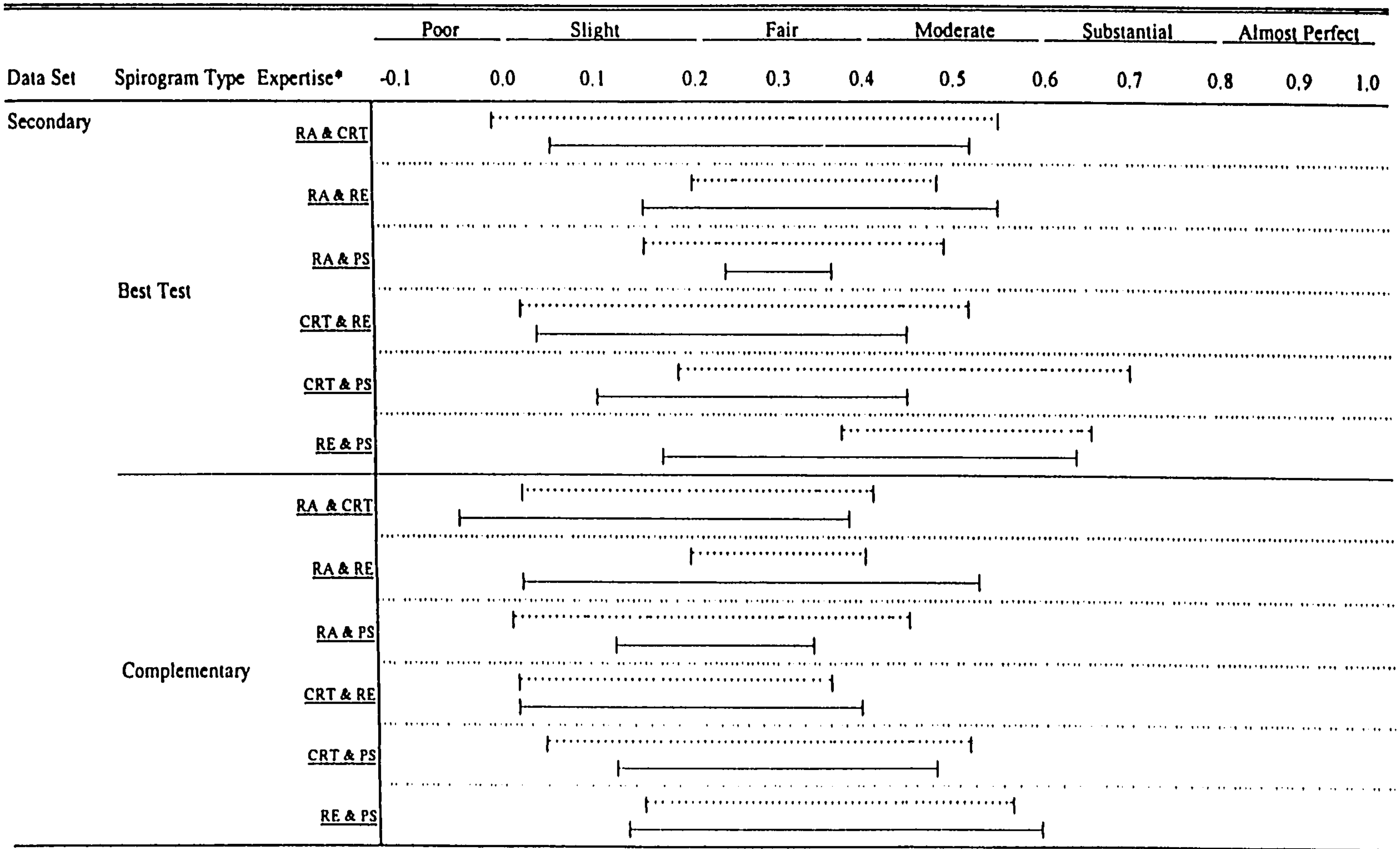


Figure 4.7 Continued



* RA = Non-Certified, Minimally Trained, Respiratory Research Assistants; CRT = Certified, Respiratory Technicians; RE = Respiratory Epidemiologists ; PS = Pulmonary Specialists

.... Based on spirograms from the minimally trained and experienced technician

____ Based on spirograms from the highly trained and experienced technician

Figure 4.8 Range of Average Kappa Values Stratified By Technician Expertise Between Expertise Levels for Both Spirogram Types in the Primary and Secondary Data Sets

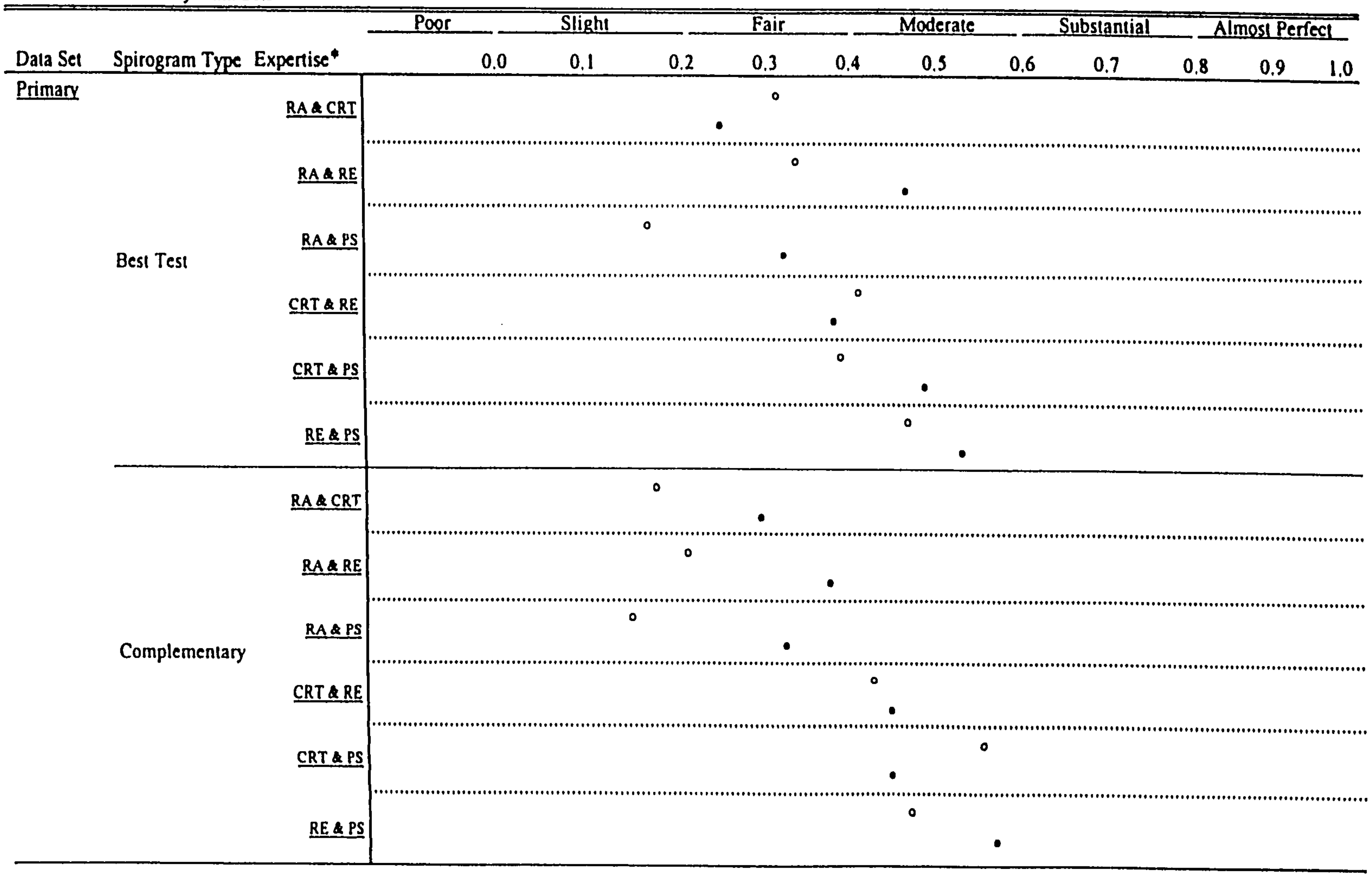
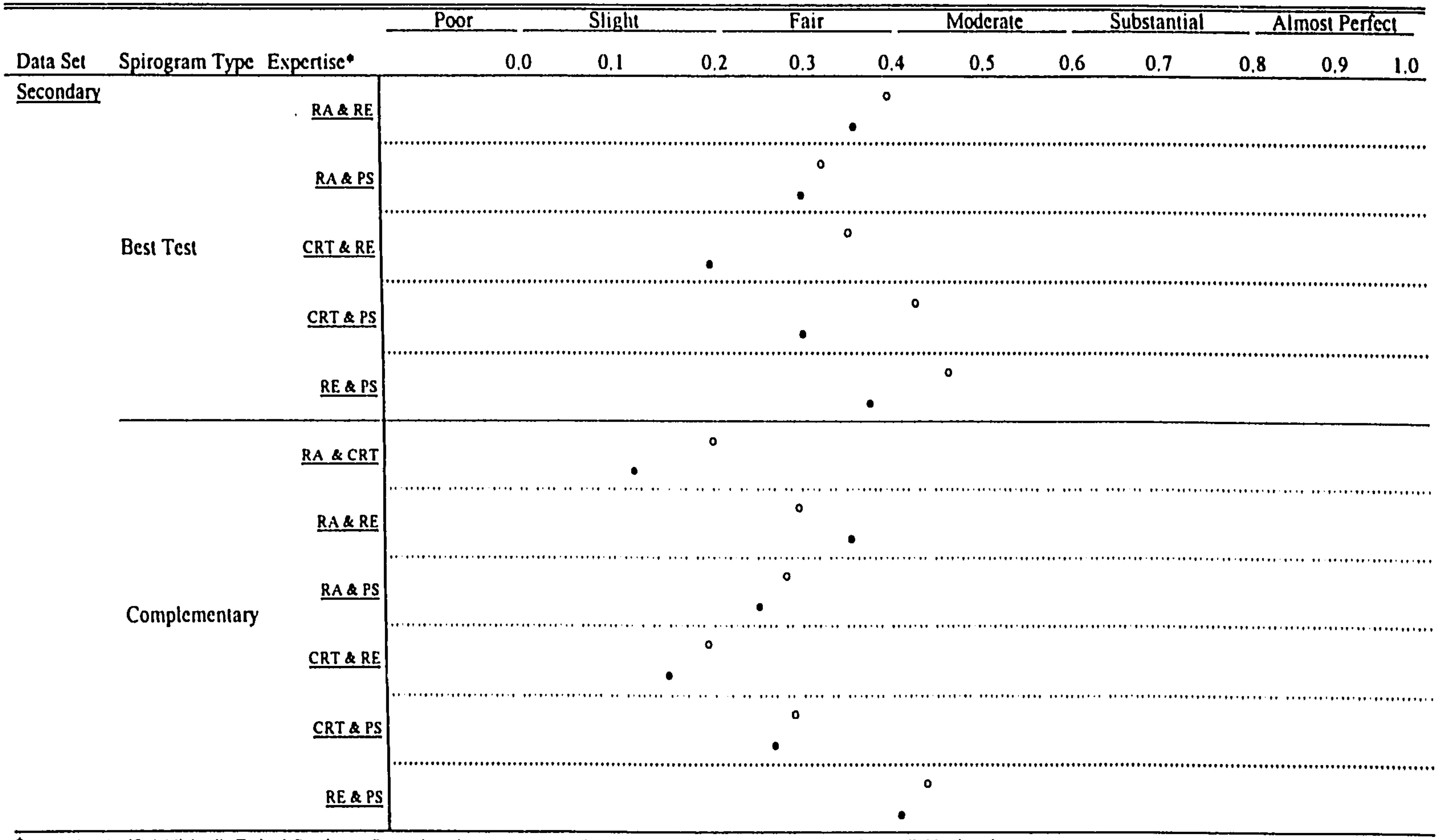


Figure 4.8 Continued



* RA = Non-Certified, Minimally Trained, Respiratory Research Assistants; CRT = Certified, Respiratory Technicians; RE = Respiratory Epidemiologists ; PS = Pulmonary Specialists

o = Based on spirograms from the minimally trained and experienced technician

• = Based on spirograms from the highly trained and experienced technician

CHAPTER FIVE

RESULTS -FACTOR ANALYSIS OF KAPPA VALUES

In this chapter, patterns of agreement among raters, identified using Principal Components Analysis techniques, are examined.

5.1 Introduction to Factor Analysis (Principal Components Analysis) Techniques

In general, factor analysis attempts to identify a small number of underlying factors, either hypothetical constructs or observable phenomena, that explain patterns of correlations among several interrelated variables (e.g., questionnaire items or response scores) (Norman and Streiner, 1997). Thus, it is frequently used in exploratory data analysis, hypothesis confirmation, and variable reduction (i.e., reduction of many variables into a more manageable number) (SPSS 9.0, 1997).

5.1.1 Derivation of Factors

The initial step of any factor analysis involves construction of a correlation matrix (illustrating relationships between individual variables). From this matrix, variables are grouped into sets or "factors". Those more highly correlated with one another than with all other variables comprise a factor. Each extracted factor resembles a linear regression equation. In principal components analysis, the first factor's linear combination of weights accounts for the maximum amount of variance among scores; successive components (factors) explain consecutively smaller portions of the total variance and are independent of one another (SPSS 9.0, 1997). Each component's corresponding variance is termed an "eigenvalue". Decisions regarding the number of variables to retain are based on the magnitude of these eigenvalues. In compliance with the "eigenvalue-one" criterion, any factor with an eigenvalue of less than one, indicating that it defines no more of the variance than does any single variable, is excluded from further analyses. Weights assigned to variables for all retained factors are subsequently examined in the form of a factor loading (components) matrix. This matrix illustrates the extent to which individual variables "load on" or correlate with a particular factor. Typically, each loading with a value greater than $5.152/(N-2)^{1/2}$ is considered statistically significant (Note: N is the number of study participants) (Norman and Streiner, 1997). Variables are assigned to the factor on which they load most highly, producing an unrotated solution. Graphs, plotting variable loadings against factors, facilitate visual assessments of the relationship among variables.

To simplify the interpretation of factors comprised of "factorially complex" variables (i.e., variables that loaded on two or more factors) rotated solutions are obtained. In brief, factor rotation techniques magnify large loadings and reduce small loadings from their unrotated values, associating each variable with a minimum number of factors. The axes of graphs plotting loadings for one factor against those of another are rotated, producing a series of points that lie close to the origin of one factor and at the extreme end of another. In turn, this process forces loadings (of factorially complex variables) out of the middle range (Norman and Streiner, 1997). The resulting rotated factor matrix is then interpreted following criteria identical to those outlined for its unrotated counterpart.

Selection of an appropriate rotation method is based on whether or not resulting factors correlate with one another. Orthogonal techniques, involving rotation of axes at right angles to each other, generate uncorrelated factors. In contrast, oblique rotations, which allow axes to rotate at angles reflecting the degree of correlation among factors, produce correlated factors.

5.1.2 Application of Principal Components Analysis

The present study used principal components analysis to explore the data for "relationships" (i.e., strength of agreement) among raters who varied in respiratory expertise. For matrix construction, variables identified individual raters. Stratified or unstratified kappa values, indicating strength of agreement between raters, replaced standard correlation coefficientts[1]. Factors (components) were then extracted from kappa matrices. As described above, those retained for inclusion in the components matrix exhibited eigenvalues of at least one. In an attempt to eliminate factorially complex variables and, in turn, simplify the interpretation of derived factors, a rotated solution was obtained. Specifically, rotation involved oblique methods, allowing for correlation among factors expected to represent raters' levels of respiratory expertise. Loadings were tabulated to form a pattern matrix. Subsequent multiplication of this pattern matrix by the factor correlation matrix generated a structure matrix. Construction of three-dimensional graphs, plotting variable loadings against the first three extracted factors, provided a supplemental visual description of correlations among variables (raters).

It is important to note that rater information was limited to respiratory expertise level and familiarity with ATS Acceptability Criteria. Consequently, this study did not attempt to identify the specific, underlying construct that each extracted factor characterized. Rather, clusters, detected graphically, were examined to determine the respiratory expertise levels of raters who appeared to correlate strongly with one another.

5.2 Results

Structure matrices and 3-dimensional plots of the first three factors extracted from all kappa score matrices were inspected for clustering of raters with similar respiratory expertise.

5.2.1 Patterns of Overall Agreement Between Raters Evaluating the Acceptability of Spirograms

Assessment of overall patterns of agreement between raters involved matrices comprised

[1] Kappa scores and Pearson's correlation coefficients range in value from -1 to +1. Zero values for both types of coefficients are interpreted similarly. A Pearson correlation coefficient of zero indicates that variables are not correlated with each other. Comparably, a zero kappa score signifies no agreement between raters beyond chance. Therefore, despite potential variations in the distribution of scores derived from each statistic, the decision to substitute Kappa for Pearson's *r* was rationalized.

exclusively of crude (unstratified) kappa values.

i) The Primary Data Set

Results pertaining to participants' best test spirograms are displayed in Tables 5.1a and 5.1b, and Figure 5.1. Two clusters of raters significantly[1] loaded on Factor 1: the first consisted of two Non-certified, Minimally Trained, Respiratory Research Assistants while the second included three raters with professional, respiratory expertise (both Pulmonary Specialists and one Certified, Respiratory Technician). Neither cluster incorporated all raters from a single category. For Factor 2, a cluster of two Respiratory Epidemiologists was detected. Significant loadings, corresponding to one Non-certified, Minimally Trained, Respiratory Research Assistant and one Certified, Respiratory Technician, formed a bipolar pattern, indicating that respective raters were highly non-correlated with each other. With respect to Factor 3, both raters, one Certified, Respiratory Technician and one Pulmonary Specialist, comprised a single group. Therefore, with one exception, complete categories of raters did not load "collectively" on any factor. Comparable findings were exhibited for complementary spirograms (refer to Tables 5.2a and 5.2b and Figure 5.2). Factor 1 consisted of two, separate clusters of raters. Both Respiratory Epidemiologists and Pulmonary Specialists, and one Certified, Respiratory Technician, formed the first cluster while its counterpart incorporated two of the three Non-certified Respiratory Technicians. The third Non-Certified, Minimally Trained, Respiratory Research Assistant and one of the two remaining Certified, Respiratory Technicians loaded significantly on Factor 2. A single pair of loadings, corresponding to the final Certified, Respiratory Technician and one Pulmonary Specialist, was identified for Factor 3 . Thus, the observed absence of clusters consisting exclusively of raters from a single category indicated that the strength of relationships (i.e., agreement) among raters was not directly correlated with their level of respiratory expertise.

[1] With respect to the Primary Data Set, variable loadings of 0.520 or greater and -0.520 or less were statistically significant (based on the equation presented in Section 5.1.2).

ii) The Secondary Data Set

For best test spirograms in the Secondary Data Set, raters, spanning multiple expertise categories loaded significantly[1] on Factor 1 (refer to Tables 5.3a and 5.3b and Figure 5.3). The resulting group comprised both Pulmonary Specialists, one Respiratory Epidemiologist, and two Certified, Respiratory Technicians. One Certified, Respiratory Technician and one Respiratory Epidemiologist were excluded from this group, suggesting that raters sharing a common expertise level "behaved" differently. In contrast, all three Non-certified, Minimally Trained, Respiratory Research Assistants clustered together on Factor 2. Similarly, loading patterns for Factor 3 displayed a single cluster which included both Respiratory Epidemiologists and Pulmonary Specialists. Parallel analysis of complementary spirograms also grouped raters with equivalent expertise on Factor 1 (refer to Tables 5.4a and 5.4b, and Figure 5.4). Each detected cluster represented a complete expertise category (specifically, the Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians Categories). However, a diverse group of raters from all professional levels collectively loaded on Factor 2. Consequently, patterns reflecting strengthened concordance within expertise categories were inconsistent across factors. Therefore, no definite correlation between the degree of inter-rater agreement and raters' levels of respiratory expertise was established.

5.2.2 The Effect of Technician Expertise on Patterns of Agreement Between Raters Evaluating the Acceptability of Spirograms

Assessment of factors extracted from matrices of stratified kappa values detected variances in rater concordance for spirograms collected from tests administered by technicians with different levels of expertise.

[1] With respect to the Secondary Data Set, variable loadings of 0.366 or greater and -0.366 or less were statistically significant (based on the equation presented in Section 5.1.2).

i) The Primary Data Set

Patterns of rater agreement regarding the acceptability of best test spirograms submitted by the minimally trained technician are displayed in Tables 5.5a and 5.5b and Figure 5.5. Two raters, each representing a separate expertise category (specifically, the Respiratory Epidemiologist and Pulmonary Specialist categories), loaded significantly[1] and similarly on the first factor. With respect to the second and third factors, Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians were grouped together. Importantly, no detected cluster comprised all raters from a single category. Those corresponding to the first three factors extracted from spirograms produced by the highly trained technician also included raters from multiple expertise categories (refer to Tables 5.6a and 5.6b and Figure 5.6). One Certified, Respiratory Technician and one Pulmonary Specialist correlated strongly with Factor 1. Pairs combining Respiratory Epidemiologists with Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians were identified for Factors 2 and 3, respectively. Since raters within categories again loaded differently on extracted factors, no relationship between technician expertise and rater expertise was evident. These findings coincided with those of the complementary spirograms (refer to Tables 5.7a, 5.7b, 5.8a and 5.8b and Figures 5.7 and 5.8). Across both technician expertise levels, no cluster defined an entire rater category. Consequently, the degree of agreement among raters with similar respiratory expertise did not appear to strengthen with technician expertise.

ii) The Secondary Data Set

As observed in the Primary Data Set, effects of technician expertise on concordance varied across raters with equivalent respiratory expertise (refer to Tables 5.9a, 5.9b, 5.10a, and 5.10b and Figures 5.9 and 5.10). Regarding the minimally trained technician's best test spirograms, all three Non-certified, Minimally Trained, Respiratory

[1] With respect to the Primary Data Set, variable loadings of 0.744 or greater and -0.744 or less were statistically significant for analyses of stratified kappa score matrices.

Research Assistants, in addition to one Respiratory Epidemiologist, clustered together on the first factor[1]. Analysis of the highly trained technician's spirograms revealed a counterpart group comprised only of the three Non-Certified, Minimally Trained, Respiratory Research Assistants . With respect to remaining factors, cluster content also differed across technician expertise levels. Two Certified, Respiratory Technicians and one Pulmonary Specialist loaded similarly on the second factor extracted from the minimally trained technicians' spirograms. In contrast, both Respiratory Epidemiologists and one Pulmonary Specialist were grouped on the equivalent factor for the highly trained technician's tracings. Patterns corresponding to the third factor also exhibited distinct clusters of raters in each technician expertise level. These findings indicated that correlations among raters varied with technician expertise. However, since clusters from neither technician expertise level represented complete rater categories, a uniform technician effect on "agreement" between raters with equivalent expertise was not evident. Results of complementary spirograms exhibited similar trends (refer to Tables 5.11a, 5.11b, 5.12a, and 5.12b and Figures 5.11 and 5.12). Across technician expertise-based factor matrices, only one detected cluster represented a single category. This cluster, identified on the first factor extracted from the highly trained technicians' spirograms, included two Non-certified, Minimally Trained Respiratory Research Assistants. Therefore "relationships" among only raters with minimal expertise appeared to strengthen with technician expertise.

5.3 Conclusions

Findings from principal components analysis of kappa score matrices were comparable to those described in Chapter Four. Since clusters primarily incorporated raters from multiple expertise categories, no relationship between concordance and raters' level of respiratory expertise was evident. In addition, patterns across both data sets did not reflect the presence of a consistent effect of technician expertise on rater agreement at the expertise category level.

[1] With respect to the Secondary Data Set, variable loadings of 0.520 or greater and -0.520 or less were statistically significant for analyses of stratified kappa score matrices.

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Best Test Spirograms in the Primary Data Set (plotted in Figure 5.1)

Table 5.1a Pattern Matrix[a]

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.398	.853	-.005
	Rater 2	-.624	-.443	.538
	Rater 3	-.840	.131	.055
Certified Respiratory Technician	Rater 4	.197	-.623	.343
	Rater 5	.056	-.010	.950
	Rater 6	.771	-.360	-.078
Respiratory Epidemiologist	Rater 7	.196	.710	.018
	Rater 8	.039	.723	.461
Pulmonary Specialist	Rater 9	.590	.391	.354
	Rater 10	.795	-.055	.236

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 6 iterations

Table 5.1b Structure Matrix

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.317	.814	.141
	Rater 2	-.676	-.415	.477
	Rater 3	-.829	.060	.091
Certified Respiratory Technicians	Rater 4	.132	-.548	.239
	Rater 5	.038	.150	.948
	Rater 6	.738	-.299	-.150
Respiratory Epidemiologists	Rater 7	.262	.731	.130
	Rater 8	.021	.793	.578
Pulmonary Specialists	Rater 9	.620	.504	.407
	Rater 10	.785	.059	.213

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Complementary Spirograms in the Primary Data set (plotted in Figure 5.2)

Table 5.2a Pattern Matrix[a]

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.072	-.937	.164
	Rater 2	-.744	.360	.281
	Rater 3	-.785	-.053	-.254
Certified Respiratory Technicians	Rater 4	.044	.899	.214
	Rater 5	-.192	.040	.949
	Rater 6	.804	.203	-.152
Respiratory Epidemiologists	Rater 7	.517	-.137	.349
	Rater 8	.625	-.236	.196
Pulmonary Specialists	Rater 9	.447	.041	.568
	Rater 10	.957	.257	-.144

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

[a] Rotation converged in 6 iterations

Table 5.2b Structure Matrix

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.033	-.928	.136
	Rater 2	-.721	.424	.150
	Rater 3	-.827	.005	-.398
Certified Respiratory Technicians	Rater 4	.012	.899	.237
	Rater 5	-.022	.070	.914
	Rater 6	.760	.136	-.001
Respiratroy Epidemiologists	Rater 7	.592	-.172	.441
	Rater 8	.680	-.283	.307
Pulmonary Specialists	Rater 9	.548	.015	.650
	Rater 10	.910	.178	.035

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Best Test Spirograms in the Secondary Data Set (plotted in Figure 5.3)

Table 5.3a Pattern Matrix [a]

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.470	.717	.060
	Rater 2	.394	.887	-.184
	Rater 3	.006	.942	-.126
Certified Respiratory Technicians	Rater 4	.020	-.642	-.629
	Rater 5	.868	.107	-.122
	Rater 6	.811	-.141	.067
Respiratory Epidemiologists	Rater 7	.151	-.046	.768
	Rater 8	-.203	-.012	.839
Pulmonary Specialists	Rater 9	.207	-.252	.798
	Rater 10	.910	.159	.145

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 5 iterations

Table 5.3b Structure Matrix

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.478	.743	.035
	Rater 2	.309	.844	.071
	Rater 3	-.064	.922	.024
Certified Respiratory Technicians	Rater 4	-.142	-.742	-.725
	Rater 5	.829	.057	.149
	Rater 6	.835	-.159	.282
Respiratory Epidemiologists	Rater 7	.378	.069	.805
	Rater 8	.043	.128	.777
Pulmonary Specialists	Rater 9	.449	-.134	.819
	Rater 10	.947	.150	.437

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Complementary Spirograms in the Secondary Data Set (plotted in Figure 5.4)

Table 5.4a Pattern Matrix [a]

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.781	-.358	.155
	Rater 2	.785	.219	.360
	Rater 3	.919	-.006	.031
Certified Respiratory Technicians	Rater 4	-.375	-.390	.489
	Rater 5	-.321	.608	.342
	Rater 6	-.606	.285	.253
Respiratory Epidemiologists	Rater 7	.502	.648	-.129
	Rater 8	-.288	-.050	-.995
Pulmonary Specialists	Rater 9	-.165	.875	-.206
	Rater 10	.018	.711	.161

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 7 iterations

Table 5.4b Structure Matrix

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.729	-.361	-.144
	Rater 2	.651	.224	.083
	Rater 3	.908	-.016	-.304
Certified Respiratory Technicians	Rater 4	-.548	-.365	.609
	Rater 5	-.454	.626	.484
	Rater 6	-.701	.303	.485
Respiratory Epidemiologists	Rater 7	.541	.636	-.285
	Rater 8	.075	-.087	-.892
Pulmonary Specialists	Rater 9	-.101	.869	-.110
	Rater 10	-.050	.718	.184

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Best test Spirograms Produced by Minimally Trained Technicians in the Primary Data Set (plotted in Figure 5.5)

Table 5.5a Pattern Matrix [a]

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.239	-.154	.872	-.210
	Rater 2	-.427	.756	.196	.014
	Rater 3	-.329	.249	.197	-.727
Certified Respiraatory Technicians	Rater 4	.072	.884	-.316	-.157
	Rater 5	.299	.912	.051	.047
	Rater 6	.333	.059	-.905	-.037
Respiratory Epidemiologists	Rater 7	.852	-.286	.093	-.145
	Rater 8	.647	.269	.534	.210
Pulmonary Specialists	Rater 9	.869	.325	-.208	-.042
	Rater 10	-.291	.029	-.013	.964

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 6 iterations

Table 5.5b Structure Matrix

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.273	.040	.853	-.222
	Rater 2	-.475	.839	.365	-.014
	Rater 3	-.417	.300	.270	-.761
Certified Respiratory Technicians	Rater 4	-.029	.793	-.083	-.109
	Rater 5	.236	.903	.295	.107
	Rater 6	.278	-.197	-.872	.034
Respiratory Epidemiologists	Rater 7	.863	-.334	.070	-.067
	Rater 8	.676	.360	.627	.269
Pulmonary Specialists	Rater 9	.829	.204	-.080	.069
	Rater 10	-.191	.080	-.056	.935

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Best Test Spirograms Produced by Highly Trained Technicians in the Primary Data Set (plotted in Figure 5.6)

Table 5.6a Pattern Matrix [a]

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified. Minimally Trained. Respiratory Research Assistants	Rater 1	-.380	.529	.381	-.321
	Rater 2	-.029	.016	.049	.957
	Rater 3	-.340	.473	.122	.635
Certified Respiratory Technicians	Rater 4	-.137	-1.072	.317	-.181
	Rater 5	.418	-.031	.839	.352
	Rater 6	.954	.106	-.349	-.007
Respiratory Epidemiologists	Rater 7	-.166	-.112	.910	-.071
	Rater 8	.085	.689	.322	-.018
Pulmonary Specialists	Rater 9	.652	.463	.118	-.326
	Rater 10	.825	-.150	.385	-.047

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 7 iterations

Table 5.6b Structure Matrix

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.404	.786	.649	-.325
	Rater 2	-.191	.030	-.098	.954
	Rater 3	-.532	.587	.216	.666
Certified Respiratory Technicians	Rater 4	.129	-.895	-.160	-.191
	Rater 5	.426	.268	.800	.149
	Rater 6	.909	-.245	-.231	-.114
Respiratory Epidemiologists	Rater 7	-.067	.343	.857	-.185
	Rater 8	-.026	.821	.649	-.094
Pulmonary Specialists	Rater 9	.624	.393	.430	-.462
	Rater 10	.890	-.135	.383	-.245

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Complementary Spirograms Produced by Minimally Trained Technicians in the Primary Data Set (plotted in Figure 5.7)

Table 5.7a Pattern Matrix[a]

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.391	-.075	.220	-.846
	Rater 2	.120	-.074	-1.001	-.061
	Rater 3	-.907	-.177	.326	.327
Certified Respiratory Technicians	Rater 4	.312	.007	.273	.773
	Rater 5	.277	.690	-.151	.346
	Rater 6	.805	-.022	.131	.261
Respiratroy Epidemiologists	Rater 7	-.272	.802	.415	-.224
	Rater 8	.273	.787	-.058	.030
Pulmonary Specialists	Rater 9	.926	.063	.105	-.037
	Rater 10	.508	-.108	.622	-.087

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 10 iterations

Table 5.7b Structure Matrix

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.344	.198	.356	-.806
	Rater 2	-.246	-.296	-.980	-.012
	Rater 3	-.833	-.465	-.032	.233
Certified Respiratory Technicians	Rater 4	.487	.150	.353	.797
	Rater 5	.534	.733	.118	.332
	Rater 6	.868	.311	.383	.344
Respiratory Epidemiologists	Rater 7	.155	.824	.552	-.322
	Rater 8	.565	.876	.246	.005
Pulmonary Specialists	Rater 9	.981	.456	.429	.054
	Rater 10	.662	.267	.763	-.045

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Complementary Spirograms Produced by Highly Trained Technicians in the Primary Data Set (plotted in Figure 5.8)

Table 5.8a Pattern Matrix[a]

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified, Minimally Trained, Research Assistants	Rater 1	.006	1.008	-.216	-.495
	Rater 2	.240	-.269	.923	-.064
	Rater 3	-.191	.263	.778	-.023
Certified Respiratory Technicians	Rater 4	.074	-.894	-.133	-.154
	Rater 5	.863	-.148	.229	-.162
	Rater 6	-.167	-.224	-.058	.975
Respiratory Epidemiologists	Rater 7	.822	.026	.054	-.437
	Rater 8	.666	.459	.028	.405
Pulmonary Specialists	Rater 9	.590	.207	-.123	.241
	Rater 10	.617	-.366	-.324	.026

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 16 iterations

Table 5.8b Structure Matrix

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.028	.849	-.108	-.178
	Rater 2	.009	-.250	.847	-.225
	Rater 3	-.425	.319	.849	-.024
Certified Respiratory Technicians	Rater 4	.178	-.954	-.194	-.402
	Rater 5	.808	-.251	.000	-.206
	Rater 6	-.109	.072	-.120	.910
Respiratory Epidemiologists	Rater 7	.795	-.165	-.127	-.414
	Rater 8	.631	.527	-.164	.553
Pulmonary Specialists	Rater 9	.613	.223	-.294	.328
	Rater 10	.736	-.428	-.518	-.036

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Best Test Spirograms Produced by Minimally Trained Technicians in the Secondary Data Set (plotted in Figure 5.9)

Table 5.9a Pattern Matrix[a]

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.646	-.626	-.046
	Rater 2	.762	.366	-.244
	Rater 3	.803	.101	-.296
Certified Respiratory Technicians	Rater 4	-.948	.093	-.306
	Rater 5	.028	.555	.333
	Rater 6	.098	.901	.081
Respiratory Epidemiologists	Rater 7	.064	.123	.874
	Rater 8	.671	.027	.261
Pulmonary Specialists	Rater 9	.014	.017	.919
	Rater 10	.110	.865	-.070

Extraction Method: Principal component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 6 iterations

Table 5.9b Structure Matrix

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.564	-.549	-.210
	Rater 2	.826	.419	-.213
	Rater 3	.834	.149	-.321
Certified Respiratory Technicians	Rater 4	-.918	-.098	-.233
	Rater 5	.084	.627	.444
	Rater 6	.215	.930	.258
Respiratory Epidemiologists	Rater 7	.031	.309	.895
	Rater 8	.660	.170	.228
Pulmonary Specialists	Rater 9	-.036	.205	.921
	Rater 10	.230	.866	.100

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Best Test Spirograms Produced by Highly Trained Technicians in the Secondary Data Set (plotted in Figure 5.10)

Table 5.10a Pattern Matrix[a]

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistant	Rater 1	.960	-.016	-.137
	Rater 2	.611	.020	.450
	Rater 3	.911	-.001	.116
Certified Respiratory Technician	Rater 4	-.225	-.772	-.117
	Rater 5	-.307	.027	.969
	Rater 6	-.730	.001	.367
Respiratory Epidemiologist	Rater 7	.303	.763	.092
	Rater 8	-.184	.787	-.391
Pulmonary Specialist	Rate 9	-.251	.885	.127
	Rater 10	.061	.060	.866

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 6 iterations

Table 5.10b Structure Matrix

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	.920	.133	.127
	Rater 2	.737	.109	.617
	Rater 3	.943	.137	.365
Certified Respiratory Technicians	Rater 4	-.375	-.805	-.170
	Rater 5	-.037	-.030	.884
	Rater 6	-.629	-.115	.167
Respiratory Epidemiologists	Rater 7	.446	.809	.167
	Rater 8	-.170	.763	-.450
Pulmonary Specialists	Rater 9	-.080	.845	.048
	Rater 10	.307	.060	.882

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Complementary Spirograms Produced by Minimally Trained Technicians in the Secondary Data Set (plotted in Figure 5.11)

Table 5.11a Pattern Matrix[a]

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified, Minimally Trained, Respiratory Research Assistant	Rater 1	.006	1.008	-.216	-.495
	Rater 2	.240	-.269	.923	-.064
	Rater 3	-.191	.263	.778	-.023
Certified Respiratory Technician	Rater 4	.074	-.894	-.133	-.154
	Rater 5	.863	-.148	.229	-.162
	Rater 6	-.167	-.224	-.058	.975
Respiratory Epidemiologist	Rater 7	.822	.026	.054	-.437
	Rater 8	.666	.459	.028	.405
Pulmonary Specialist	Rater 9	.590	.207	-.123	.241
	Rater 10	.617	-.366	-.324	.026

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normatlization
[a] Rotation converged in 16 iterations

Table 5.11b Structure Matrix

Respiratory Expertise Category	Variable	Factor			
		1	2	3	4
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.028	.849	-.108	-.178
	Rater 2	.009	-.250	.847	-.225
	Rater 3	-.425	.319	.849	-.024
Certified Respiratory Technicians	Rater 4	.178	-.954	-.194	-.402
	Rater 5	.808	-.251	.000	-.206
	Rater 6	-.109	.072	-.120	.910
Respiratory Epidemiologists	Rater 7	.795	-.165	-.127	-.414
	Rater 8	.631	.527	-.164	.553
Pulmonary Specialists	Rater 9	.613	.223	-.294	.328
	Rater 10	.736	-.428	-.518	-.036

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

Results of Principal Components Analysis of Inter-rater Agreement Regarding the Acceptability of Complementary Spirograms Produced by Highly Trained Technicians in the Secondary Data Set (plotted in Figure 5.12)

Table 5.12a Pattern Matrix[a]

Respiratory Expertise Category	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.752	.024	.111
	Rater 2	-.028	.063	.838
	Rater 3	-.553	.338	.577
Certified Respiratory Technicians	Rater 4	.098	-.874	.030
	Rater 5	.769	-.030	.147
	Rater 6	.748	-.058	.053
Respiratory Epidemiologists	Rater 7	.074	.742	.183
	Rater 8	-.245	.580	-.640
Pulmonary Specialists	Rater 9	.682	.598	-.049
	Rater 10	.377	.269	.470

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization
[a] Rotation converged in 5 iterations

Table 5.12b Structure Matrix

Respiratory Expertise Catergory	Variable	Factor		
		1	2	3
Non-certified, Minimally Trained, Respiratory Research Assistants	Rater 1	-.733	.044	-.019
	Rater 2	.120	.200	.843
	Rater 3	-.452	.433	.534
Certified Respiratory Technicians	Rater 4	.105	-.869	-.095
	Rater 5	.795	-.007	.279
	Rater 6	.758	-.051	.176
Respiratory Epidemiologists	Rater 7	.105	.772	.317
	Rater 8	-.359	.476	-.589
Pulmonary Specialists	Rater 9	.673	.589	.170
	Rater 10	.460	.345	.581

Extraction Method: Principal Component Analysis
Rotation Method: Promax with Kaiser Normalization

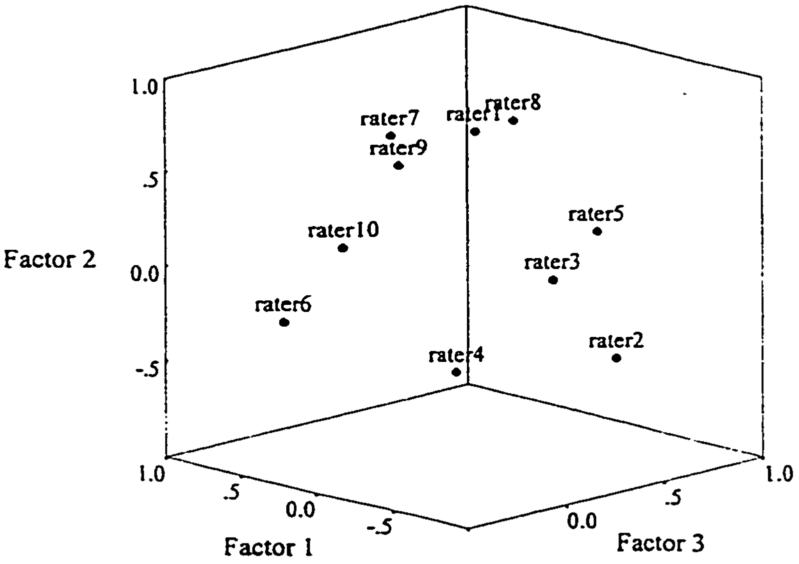


Figure 5.1 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of best test spirograms in the Primary Data Set

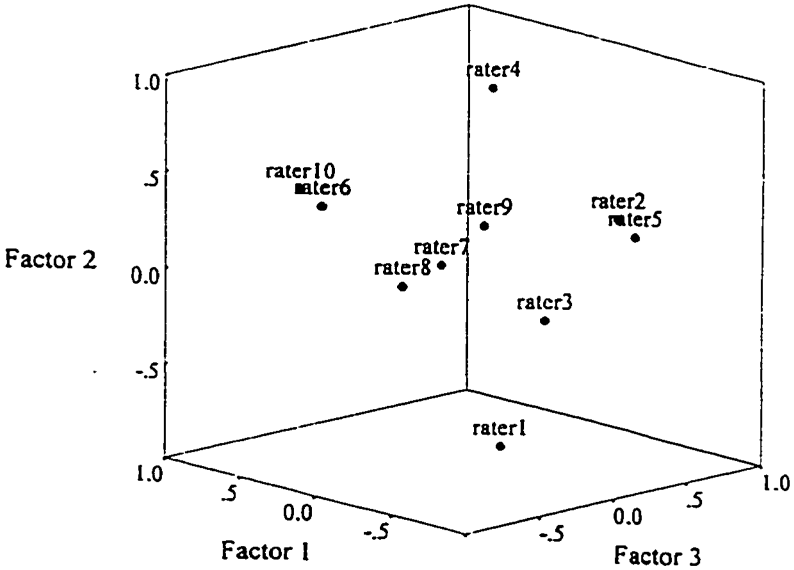


Figure 5.2 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of complementary spirograms in the Primary Data Set

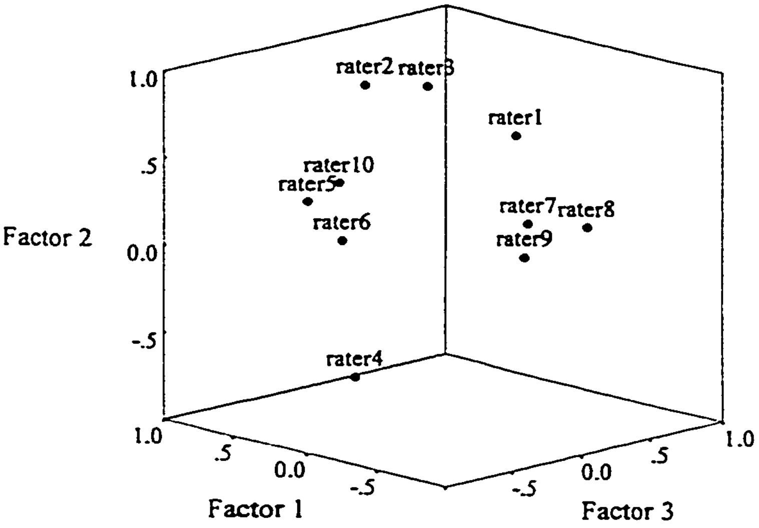


Figure 5.3 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of best test spirograms in the Secondary Data Set

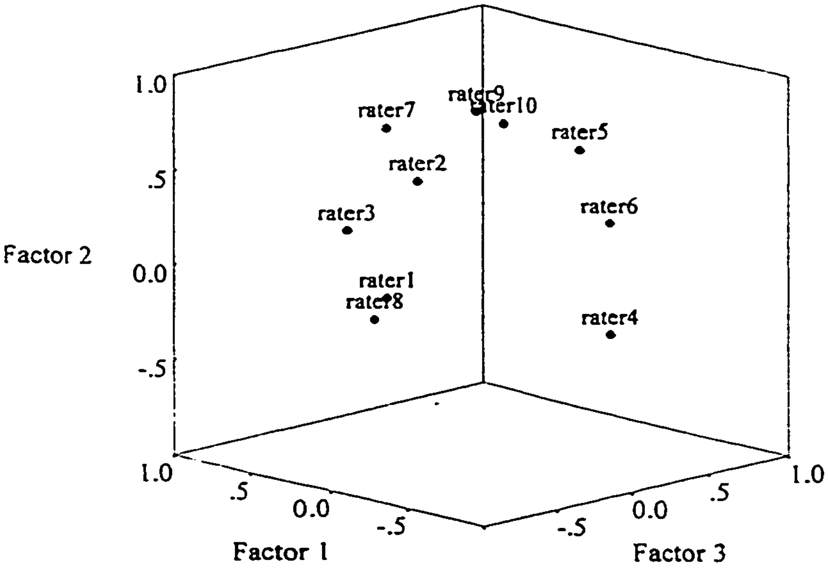


Figure 5.4 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of complementary spirograms in the Secondary Data Set

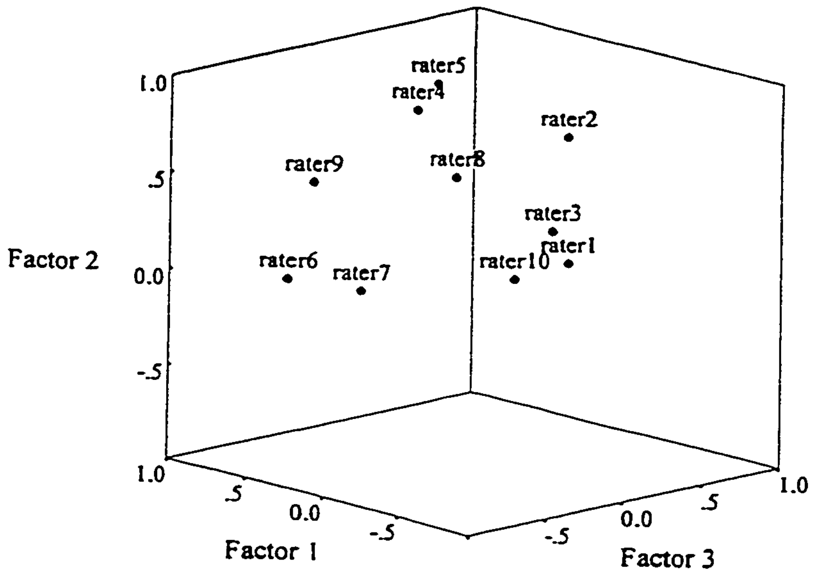


Figure 5.5 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of best test spirograms produced by the minimally trained technician in the Primary Data Set

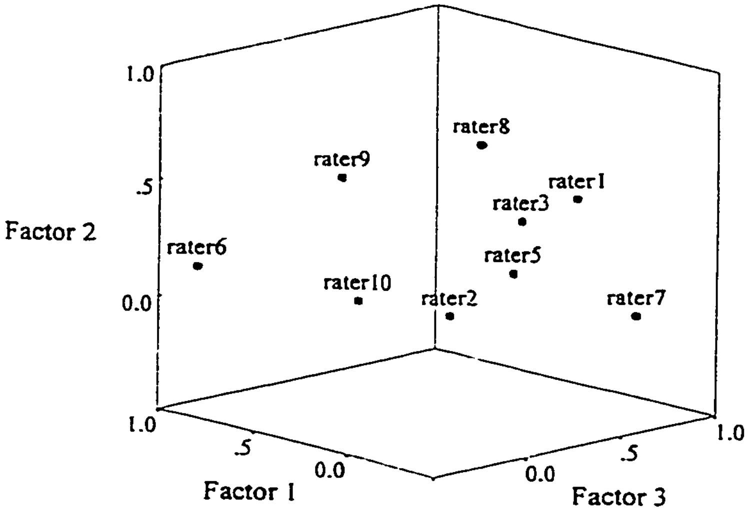


Figure 5.6 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of best test spirograms produced by the highly trained technician in the Primary Data Set

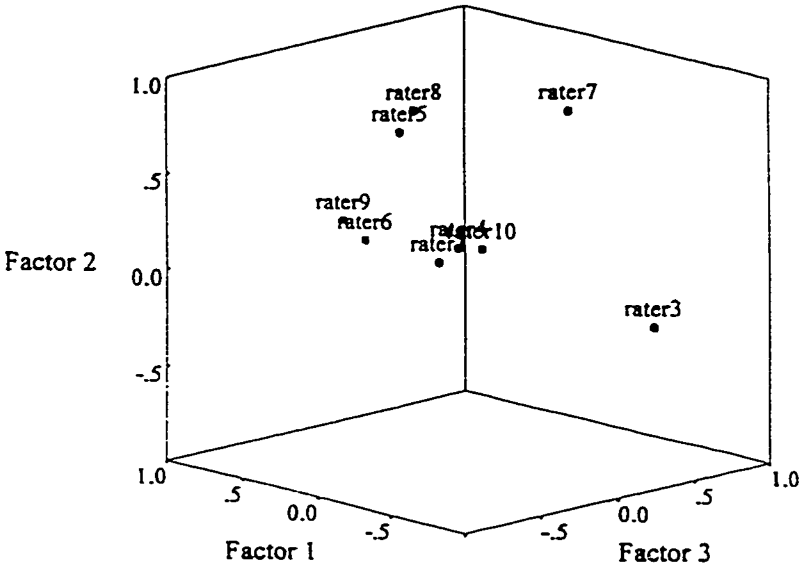


Figure 5.7 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of complementary spirograms produced by the minimally trained technician in the Primary Data Set

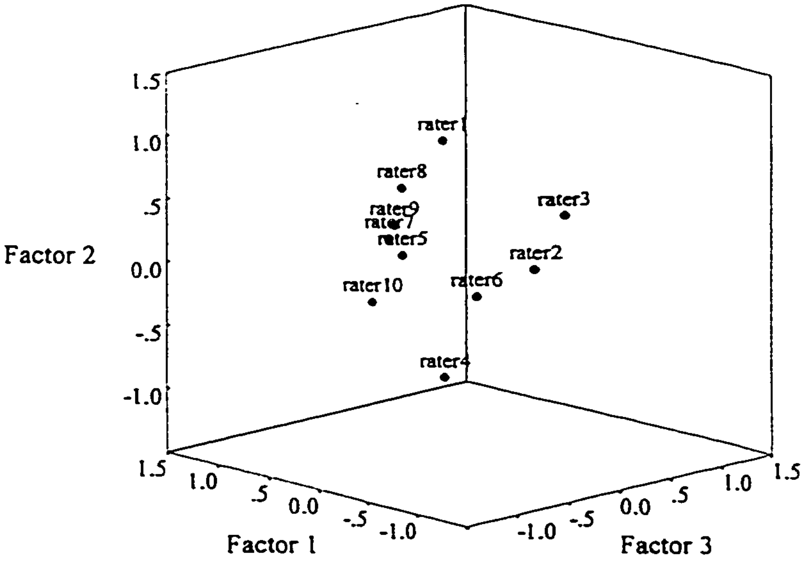


Figure 5.8 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of complementary spirograms produced by the highly trained technician in the Primary Data Set

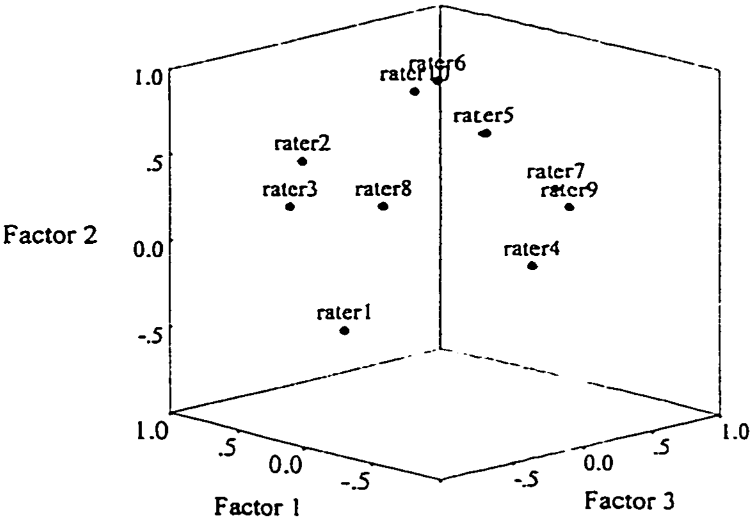


Figure 5.9 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of best test spirograms produced by minimally trained technicians in the Secondary Data Set

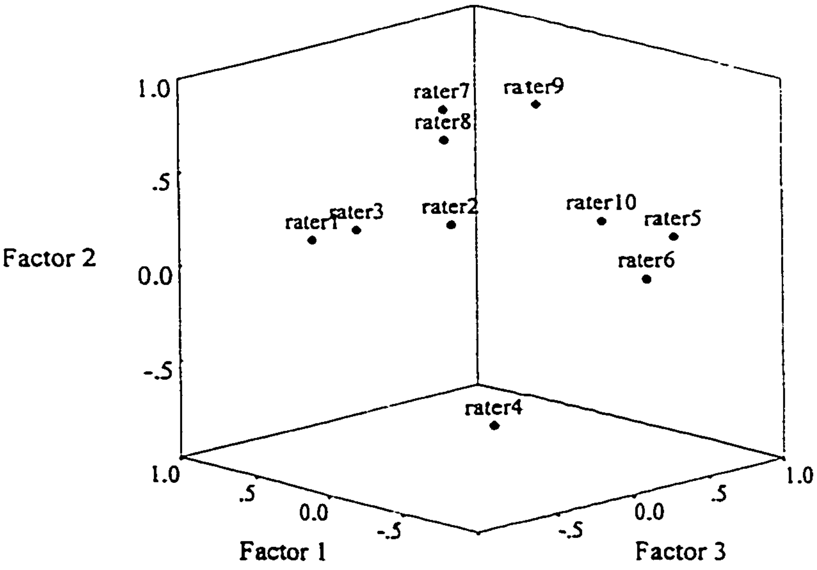


Figure 5.10 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of best test spirograms produced by the highly trained technician in the Secondary Data Set

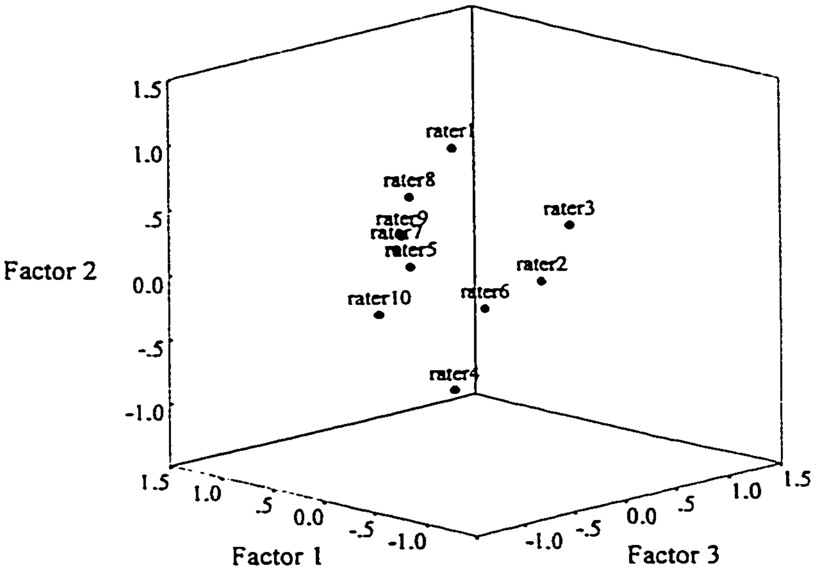


Figure 5.11 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of complementary spirograms produced by minimally trained technicians in the Secondary Data Set

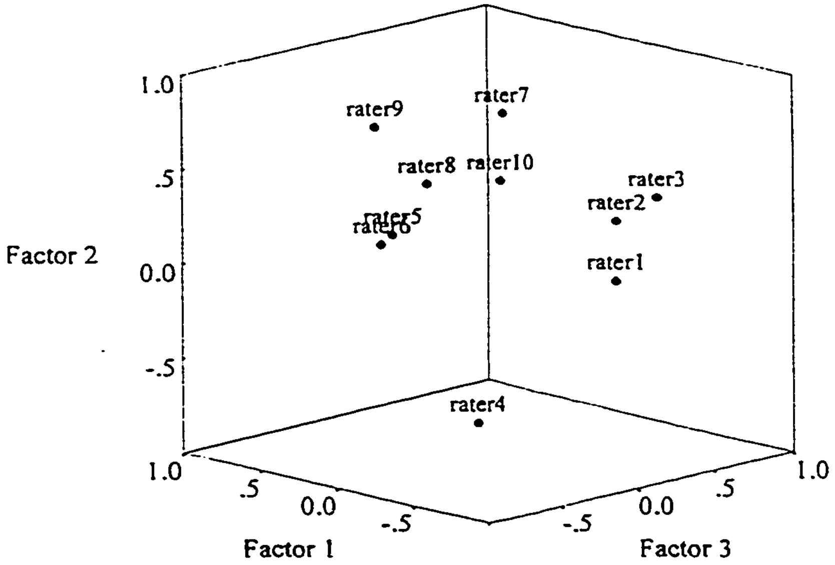


Figure 5.12 Rotated three dimensional plot of factors extracted from principal component analysis of inter-rater agreement regarding the acceptability of complementary spirograms produced by the highly trained technician in the Secondary Data Set

CHAPTER SIX

RESULTS -THE ACCEPTABILITY OF SPIROMETRIC RESULTS

This chapter examines variations in the proportion of acceptable spirometric tests conducted by minimally trained or highly trained technicians according to raters' expertise.

6.1 Determination of Acceptable Spirometric Results

To determine whether spirometric test acceptability (as opposed to strength of agreement) varied with rater expertise, a comparative analysis of the proportion of spirograms interpreted as acceptable by *all* raters within each category was completed. In addition, tracings, assembled into technician expertise-based subsets, facilitated an assessment of technician effects on raters' evaluations. Differences were subsequently examined for statistical significance using chi-square tests.

Three, separate, comparative analyses were performed, the first of which examined best test spirograms. The second and third analyses included both best test and complementary spirograms (i.e., spirogram pairs), thereby quantifying differences in the acceptability of complete test sessions. However, while one considered all spirogram pairs, the second involved only those complying with ATS reproducibility criteria (ATS, 1994). Respiratory epidemiologic studies, in the past, have frequently extracted lung function data from each of these three sources (McKay, 1991).

6.2 Results

6.2.1 Acceptability of Best Test Spirograms

i) The Primary Data Set

a) Overall Acceptability

The proportion of best test spirograms interpreted as acceptable by each category of raters is presented in Table 6.1. Across Non-certified, Minimally Trained, Respiratory

Research Assistants and Certified, Respiratory Technicians, approximately one-third of tracings received an acceptable rating (35.0% and 32.0%, respectively). Although a slightly higher value was calculated for both Respiratory Epidemiologists and Pulmonary Specialists (each exhibited a value of 45.0 %), differences were statistically non-significant ($0.05 < \text{p-value} < 0.10$). Thus, these results did not indicate a relationship between the proportion of spirograms deemed acceptable and rater's respiratory expertise.

b) Effects of Technician Expertise on Acceptability

With respect to each rater category, equal proportions of best test spirograms produced by minimally trained and highly trained technicians were evaluated as acceptable (p-values > 0.2) (refer to Table 6.1). Across Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians, one-third of tracings from each technician received an "acceptable" rating. While a slightly higher proportion (two-fifths) was acceptable according to Respiratory Epidemiologists and Pulmonary Specialists, differences between rater categories were not statistically significant ($0.05 < \text{p-value} < 0.10$). Thus, regardless of rater expertise, no evidence of a technician effect on the assessment of best test spirogram acceptability was presented.

During the assembly of spirogram sets, a distinct, graphical anomaly was noticed on several spirometric records (refer to Figure 6.1). Although not determined conclusively, its origin was thought to be equipment-related. To ensure that records displaying this artifact were not distorting a possible technician effect on spirogram acceptability, the analysis was repeated following their exclusion (Note: Approximately 34.0% and 24.6% of spirograms from the minimally trained and highly trained technicians, respectively, were excluded). Results, summarized in Table 6.1a, generated analogous trends. For each rater category, a parallel increase in the proportion of acceptable best test spirograms from the two levels of technicians occurred (p-values > 0.4). Further, between rater categories, values corresponding to each technician also varied nonsignificantly. Thus, consistent with the above findings, no effect of technician expertise on raters' evaluations of spirogram acceptability was found.

ii) The Secondary Data Set

a) Overall Acceptability

Across levels of rater expertise, statistically significant differences in the proportion of acceptable best test spirograms were noted (p-value<0.01) (refer to Table 6.2). While categories comprised of Non-certified, Minimally Trained, Respiratory Research Assistants, Respiratory Epidemiologists, and Pulmonary Specialists each evaluated at least half of the tracings as acceptable (50.0%, 66.0%, and 55.0%, respectively), the Certified, Respiratory Technician group assigned "acceptable" ratings to a substantially lower proportion (18.0%). These findings, indicating that the proportion of acceptable spirograms was, to a degree, associated with rater expertise, contrasted those of the Primary Data Set.

b) Effects of Technician Expertise on Acceptability

Differences in the proportion of acceptable best test spirograms between technician expertise levels varied with rater expertise (refer to Table 6.2). Of tracings evaluated as acceptable by Non-certified, Minimally Trained, Respiratory Research Assistants, approximately half corresponded to spirometry sessions conducted by the highly trained technician. Consequently, no technician effect was observed. However, across categories of Certified, Respiratory Technicians, Respiratory Epidemiologists, and Pulmonary Specialists, the proportion of acceptable spirograms from the highly trained technician exceeded that of the minimally trained counterpart by almost two-fold. Thus, contrary to findings from the Primary Data Set, a uniform effect of technician expertise on spirogram acceptability was evident among raters with professional levels of respiratory expertise.

6.2.2 Acceptability of Spirometric Test Sessions

i) The Primary Data Set

a) Overall Acceptability

The proportion of spirometric test sessions (i.e., pairs of best test and complementary spirograms) classified as acceptable according to each category of raters is presented in Table 6.3. All values were approximately 10% smaller than those derived from best test spirograms. Therefore, differences in the proportion of acceptable test sessions across categories remained statistically non-significant [Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians: ≈25% test session acceptability, Respiratory Epidemiologists and Pulmonary Specialists:≈38% test session acceptability (p-value > 0.05)]. In turn, no relationship between the proportion of test sessions deemed acceptable and rater expertise was detected.

b) Effects of Technician Expertise on Acceptability

Regardless of rater category, the proportions of acceptable test sessions corresponding to technician expertise levels did not differ (all p-values > 0.50) (refer to Table 6.3). Thus, trends coinciding with those established for best test spirograms suggested the absence of a "technician effect" on test session acceptability.

ii) The Secondary Data Set

a) Overall Acceptability

Within each rater category, the proportion of acceptable spirometry sessions was less than the proportion of acceptable best test spirograms (refer to Table 6.4). In addition, values varied statistically significantly across categories (p-value < 0.01). While two-thirds of test sessions were considered acceptable by Respiratory Epidemiologists, only one-third received acceptable ratings among Non-certified, Minimally Trained, Respiratory Research Assistants and Pulmonary Specialists. Moreover, a substantially lower proportion (one-sixth) was acceptable according to the Certified, Respiratory Technicians. Thus, despite variations in values, neither a positive nor negative correlation between test session acceptability and rater expertise was established.

b) Effects of Technician Expertise on Acceptability

In categories of raters with professional, respiratory expertise, the proportion of acceptable test sessions significantly increased with technician expertise (refer to Table 6.4). Among Certified, Respiratory Technicians, Respiratory Epidemiologists and Pulmonary Specialists, values pertaining to the highly trained technician were a minimum of 1.5 times higher than those associated with minimally trained technicians. Similar to trends established for best test spirograms, technician expertise did not appear to influence the proportion of sessions evaluated as acceptable by Non-certified, Minimally Trained, Respiratory Research Assistants. Equal values were exhibited for both minimally trained and highly trained technicians.

6.2.3 Acceptability of ATS Reproducible Spirometric Test Sessions

i) The Primary Data Set

Of 100 spirometry sessions, 62 satisfied ATS reproducibility criteria (refer to Section 1.4) and were subsequently included in this analysis.

a) Overall Acceptability

Results, identical to those of best test spirograms, demonstrated no statistically significant differences in the proportion of both acceptable and reproducible spirometry sessions across rater categories (p-value > 0.05). Again, one-third of the sessions were considered acceptable according to Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians. A similar proportion (two-fifths) received an acceptable rating by Respiratory Epidemiologists and Pulmonary Specialists. Therefore, no indication of a relationship between the acceptability of reproducible test sessions and raters' respiratory expertise was detected.

b) Effects of Technician Expertise on Acceptability

Approximately three-fifths of spirometry sessions conducted by each technician met ATS reproducibility criteria, suggesting no effect of technician expertise on test session reproducibility (p-value = 0.44) (refer to Table 6.5). Proportions of acceptable, reproducible spirometry sessions from both technician expertise levels were also compared (refer to Table 6.5). Results exhibited no statistically significant differences between technician types, regardless of rater category (all p-values > 0.2). Thus, technician expertise did not appear to affect the acceptability of ATS-reproducible test sessions.

ii) The Secondary Data Set

Two-thirds of spirometric test sessions, a proportion comparable to the Primary Data Set, satisfied ATS reproducibility criteria.

a) Overall Acceptability

Coinciding with results of both best test spirograms and spirometry sessions, the proportion of acceptable, reproducible sessions differed significantly between rater categories (p-value < 0.01) (refer to Table 6.6). While Certified, Respiratory Technicians considered one-tenth of sessions acceptable, raters in each of the remaining categories judged a substantially greater proportion acceptable (values were at least four times higher than that corresponding to Certified, Respiratory Technicians). However, since this variation did not distinguish raters with minimal, respiratory expertise from those with professional levels, a relationship between the acceptability of reproducible test sessions and raters' levels of respiratory expertise was not evident.

b) Effects of Technician Expertise on Acceptability

Contrary to those of the Primary Data Set, a statistically significant difference in the percentage of ATS-reproducible test sessions was observed across technician expertise levels (refer to Table 6.6). Values of 55.0 % and 80.0 %, which corresponded to the minimally trained technician and highly trained technician, respectively, indicated a positive association between test session reproducibility and technician expertise. However, no evidence of a technician effect on raters' assessments of reproducible test session acceptability was detected. Regardless of rater category, differences in the proportion of reproducible test sessions evaluated as acceptable across technician expertise levels were not statistically significant (all p-values > 0.15). Therefore, technician expertise appeared to affect only the reproducibility of spirometry sessions.

6.3 Conclusions

No indication of a relationship between the proportion of acceptable best test spirograms, complete spirometry sessions, or ATS-reproducible tests and raters' levels of respiratory expertise was evident in the Primary Data Set. Further, the expertise level of technicians directing spirometric tests did not appear to influence raters' interpretations.

Results of the Secondary Data Set exhibited different findings. The proportion of acceptable best test spirograms and spirometry sessions (non-reproducible and reproducible) varied across rater categories. However, such variances did not correlate with raters' degree of respiratory expertise. In contrast, with regard to technician expertise, a positive effect on the proportion of both acceptable best test spirograms and complete spirometry sessions was noted for raters with professional, respiratory expertise.

Therefore, the only comparable finding between data sets was the absence of a technician effect on the acceptability of ATS-reproducible sessions.

Table 6.1 Proportion of Best Test Spirograms Interpreted as Acceptable From Participants' First Spirometric Test Sessions in the Primary Data Set

		Level of Technician Conducting Spirometric Test		
Acceptability of Best Test Spirograms According Rater Category	Acceptable Best Test Spirograms n (%)*	Minimally Trained Technician n (%)	Highly Trained Technician n (%)	p-value•
Non-certified, Minimally Trained, Respiratory Research Assistants	35 (35.0)	-----	-----	0.542
		15 (31.9)	20 (37.7)	
Certified Respiratory Technicians	32 (32.0)	-----	-----	
		17 (36.2)	15 (28.3)	0.400
Respiratory Epidemiologists	45 (45.0)	-----	-----	
		20 (42.6)	25 (47.2)	0.643
Pulmonary Specialists	45 (45.0)	-----	-----	
		24 (51.1)	21 (39.6)	0.251
Total Number of Best Test Spirograms		47	53	

* Number and percent of acceptable best test spirograms
• p values based on chi-square test statistic, unless otherwise indicated

Table 6.1a Proportion of Best Test Spirograms Interpreted As Acceptable From Participants' First Spirometric Test Sessions in the Primary Data Set after Exclusion of Those Displaying a Possible Equipment-Related Artifact on the Graphic Record

		Level of Technician Conducting Spirometric Test		
Acceptability of Best Test Spirograms According To Rater Category	Acceptable Best Test Spirograms n (%)*	Minimally Trained Technician n (%)	Highly Trained Technician n (%)	p-value•
Non-certified, Minimally Trained Respiratory Research Assistants	35 (49.3)	-----	-----	
		14 (45.2)	21 (52.5)	0.811
Certified Respiratory Technicians	26 (36.6)	-----	-----	
		13 (41.9)	13 (32.5)	0.463
Respiratory Epidemiologists	39 (55.0)	-----	-----	
		16 (51.6)	23 (57.5)	0.639
Pulmonary Specialists	36 (50.8)	-----	-----	
		17 (54.8)	19 (47.5)	0.634
Total Number of Remaining Best Test Spirograms		31	40	

* Number and percent of acceptable best test spirograms
• p-values based on chi-square test statistic, unless otherwise indicated

Table 6.2 Proportion of Best Test Spirograms Interpreted as Acceptable in the Secondary Data Set

		Level of Technician Conducting Spirometric Test		
Acceptability of Best Test Spirograms According Rater Category	Acceptable Best Test Spirograms n (%)	Minimally Trained Technician n (%)	Highly Trained Technician n (%)	p-value[:]
Non-certified, Minimally Trained Respiratory Research Assistants	100 (50.0)	-----	-----	
		45 (45.0)	55 (55.0)	0.157
Certified Respiratory Technicians	36 (18.0)	-----	-----	
		12 (12.0)	24 (24.0)	<0.0001
Respiratory Epidemiologists	132 (66.0)	-----	-----	
		55 (55.0)	83 (83.0)	<0.0001
Pulmonary Specialists	106 (55.0)	-----	-----	
		39 (39.0)	67 (67.0)	<0.0001
Total Number of Best Test Spirograms		100	100	

[*] Number and percent of acceptable best test spirograms

[:] p-values based upon chi-square test statistic, unless otherwise indicated

Table 6.3 Proportion of First Spirometric Test Sessions Interpreted as Acceptable in the Primary Data Set

		Level of Technician Conducting Spirometric Test		
Acceptability of Spirometric Test Session According Rater Category	Acceptable Test Sessions n (%)*	Minimally Trained Technician n (%)	Highly Trained Technician n (%)	p-value:
Non-certified, Minimally Trained, Respiratory Research Assistants	24 (24.0)	-----	-----	
		10 (21.3)	14 (26.4)	0.548
Certified Respiratory Technicians	26 (26.0)	-----	-----	
		13 (27.7)	13 (24.5)	0.722
Respiratory Epidemiologists	39 (39.0)	-----	-----	
		17 (36.2)	22 (41.5)	0.585
Pulmonary Specialists	37 (37.0)	-----	-----	
		17 (36.2)	20 (37.7)	0.871
Total Number of Test Sessions	100	47	53	

*Number and percent of acceptable first spirometric test sessions
: p-values based on chi-square test statistic, unless otherwise indicated

Table 6.4 Proportion of Spirometric Test Sessions Interpreted as Acceptable in the Secondary Data Set

		Level of Technician Conducting Spirometric Test		
Acceptability of Spirometric Test Session According To Rater Category	Acceptable Test Sessions n (%)*	Minimally Trained Technician n(%)	Highly Trained Technician n(%)	p-value:
Non-certified, Minimally Trained, Respiratory Research Assistants	71 (35.5)	-----	-----	
		33 (33.0)	38 (38.0)	0.460
Certified Respiratory Technicians	32 (16.0)	-----	-----	
		11 (11.0)	21 (21.0)	0.054
Respiratory Epidemiologists	117 (58.5)	-----	-----	
		49 (49.0)	68 (68.0)	0.006
Pulmonary Specialists	69 (34.5)	-----	-----	
		27 (27.0)	42 (42.0)	0.026
Total Number of Test Sessions		100	100	

*Number and percent of acceptable first spirometric test sessions
: p-values based on chi-square test statistic, unless otherwise indicated

Table 6.5 Proportion of ATS-Reproducible First Spirometric Test Sessions Interpreted as Acceptable in the Primary Data Set

		Level of Technician Conducting Spirometric Test		
Acceptability of Reproducible Test Session According To Selected Rater Categories	Acceptable, Reproducible Test Sessions n (%)*	Minimally Trained Technician n (%)	Highly Trained Technician n (%)	p-value‡
Non-certified, Minimally Trained, Respiratory Research Assistants	21 (33.9)	-----	-----	
		9 (29.0)	12 (38.7)	0.421
Certified Respiratory Technicians	20 (32.3)	-----	-----	
		12 (38.7)	8 (25.8)	0.277
Respiratory Epidemiologists	28 (45.2)	-----	-----	
		12 (38.7)	16 (51.2)	0.444
Pulmonary Specialists	26 (42.0)	-----	-----	
		12 (38.7)	14 (45.2)	0.797†
Total Number of Reproducible Test Sessions		31 (66.0)§	31 (58.5)§	0.443

* Number and percent of acceptable "best test" spirograms

‡ p-value based on chi-square test statistic, unless otherwise indicated

† p-value based on Fisher's Exact Test statistic

§ Percent based upon total number first test sessions administered by respective technician

Table 6.6 Proportion of ATS-Reproducible Test Sessions Interpreted as Acceptable in the Secondary Data Set

		Level of Technician Conducting Spirometric Test		
Acceptability of Reproducible Test Session According To Rater Category	Acceptable, Reproducible Test Sessions n (%)*	Minimally Trained Technician n (%)	Highly Trained Technician n (%)	p-value‡
Non-certified, Minimally Trained, Respiratory Research Assistants	52 (38.5)	-----	-----	
		24 (43.6)	28 (35.0)	0.311
Certified Respiratory Technicians	14 (10.4)	-----	-----	
		5 (9.1)	9 (11.3)	0.152†
Respiratory Epidemiologists	84 (62.3)	-----	-----	
		32 (58.2)	52 (65.0)	0.422
Pulmonary Specialists	64 (47.5)	-----	-----	
		24 (43.6)	40 (50.0)	0.467
Total Number of Reproducible Test Sessions		55 (55.0)§	80 (80.0)§	<0.0001

* Number and percent of acceptable "best test" spirograms

‡ p-value based on chi-square test statistic, unless otherwise indicated

† p-value based on Fisher's Exact Test statistic

§ Percent based upon total number test sessions administered by respective technician

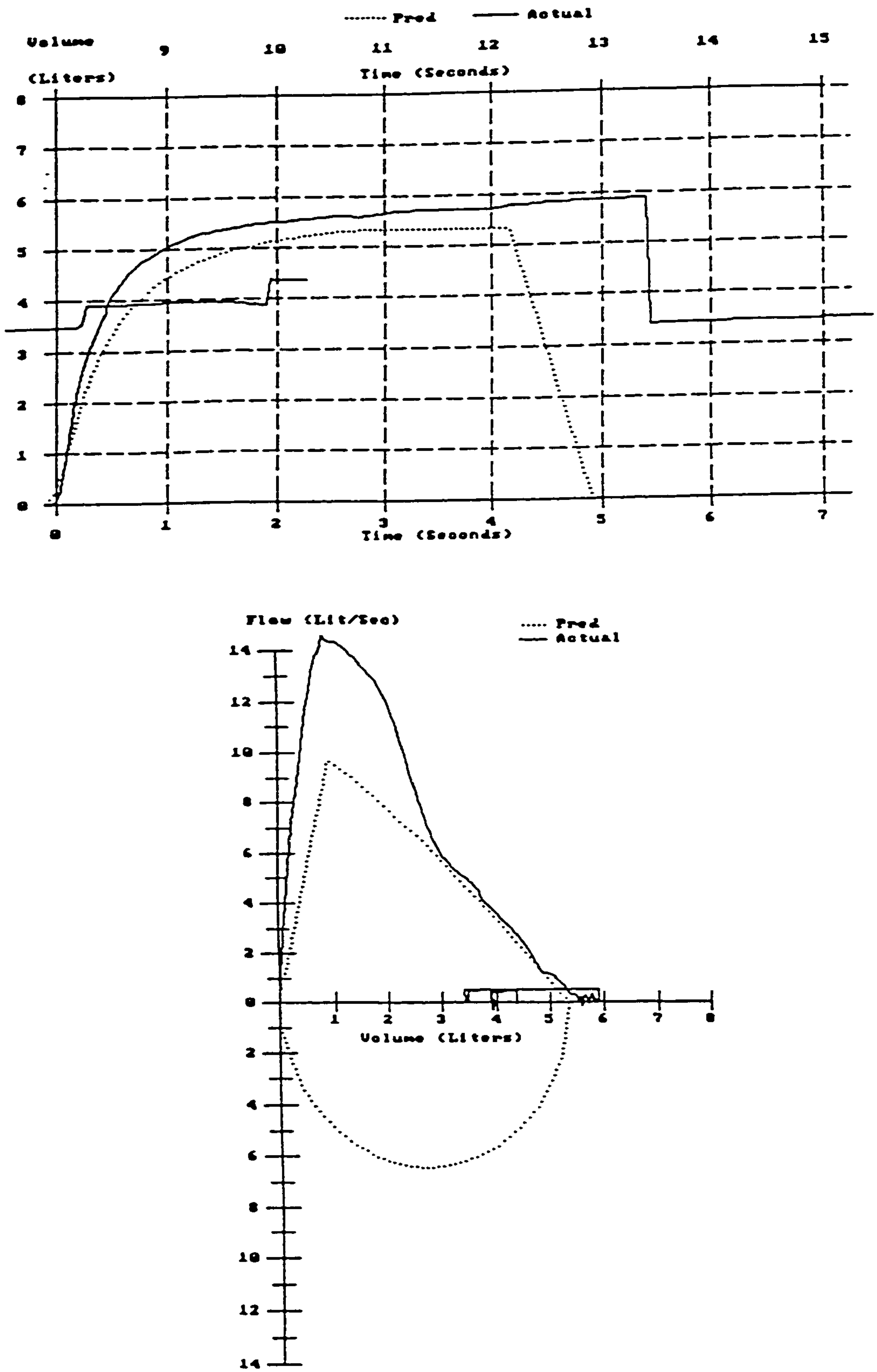


Figure 6.1 Example of the questionable artifact displayed on several spirometric tracings

CHAPTER SEVEN

RESULTS – REASONS FOR SPIROMETRIC TEST FAILURE

This chapter provides an overall analysis of the influence of raters' respiratory expertise on the interpretation of spirograms evaluated as not acceptable.

7.1 Qualitative Analysis of Spirometric Test Failure

Assessment of differences in raters' rationale for spirogram non-acceptability between respiratory expertise levels employed qualitative techniques. Briefly, comments contributed by raters were coded and collapsed into a series of thematic categories based on ATS acceptability criteria (ATS, 1994) (refer back to Section 2.11.1). Analysis, therefore, involved recognition of trends derived from these coded responses.

To determine whether tendencies to attribute spirogram non-acceptability to start-of-test failure, end-of-test failure, or both varied with raters' respiratory expertise, differences in the proportions of non-acceptable spirograms (judged accordingly by *all* raters within each category) failing to meet each set of ATS criteria (i.e., start-of-test, end-of-test, or both) were examined across levels of expertise. It was hypothesized that Non-certified, Minimally Trained, Respiratory Research Assistants, whose knowledge of spirometric curve patterns was limited to that acquired through inspection of diagrams presented within the ATS criteria document (ATS, 1994), would define spirogram non-acceptability less selectively, ascribing the majority of rejections to non-compliance with both start-of-test and end-of-test criteria. Subsequent classification of spirograms into subsets based on technician expertise permitted an examination of the effects of technician expertise on raters' interpretations. Importantly, since a different set of non-acceptable spirograms corresponded to each expertise category, variations in the *types* of overall explanations, rather than variations in comments on a single spirogram, were examined. However, an additional comparative analysis of statements extracted from spirograms rejected by all raters provided a basic indication of whether a relationship existed between respiratory expertise and interpretation of a common spirogram pattern.

All of the above analyses incorporated statements extracted exclusively from participants' best test spirograms. In previous chapters, analogous trends were observed for both best test and complementary spirograms. Thus, separate assessment of complementary spirograms was not completed.

7.2 Results - Explanations for Spirometric Test Failure

7.2.1 Variances in Explanations for Spirometric Test Failure Within Rater Categories

i) The Primary Data Set

a) Overall Non-acceptability

Evaluations from Non-certified, Minimally Trained, Respiratory Research Assistants are summarized in Table 7.1a. Fifteen percent of spirograms were judged non-acceptable by all raters. Of these tracings, 46.7% failed to comply with both ATS start-of-test and end-of-test criteria. The remaining 53.3% were rejected on recognition of an unsatisfactory test termination. Explanations for start-of-test failure included participant "hesitation" "coughing" or "variable effort". Raters applied the latter two phrases interchangeably. Curves, indicating end-of-test failure, displayed characteristics considered to reflect either "incomplete exhalation", "slight inhalation", or a "leak". In addition to these precise statements, generalized explanations were also contributed and recorded. They included "poor start", "poor termination", and "flow fluctuations".

Certified Respiratory Technicians interpreted 21.0% of spirograms as non-acceptable (refer to Table 7.1b). Slightly over 50.0% of rejected tracings failed to meet ATS start-of-test and end-of-test criteria. Those remaining had complied with only one of the two sets of criteria. Half displayed an inadequate start and half exhibited an unsatisfactory termination. Explanations for start-of-test failure were attributed to participant "coughing" or "variable effort". Once again, these statements appeared together, indicating that a single characteristic was defined differently across raters. With respect to end-of-test failure, a comment, suggesting the possibility of a technical (i.e., equipment) problem, accompanied those previously mentioned by Non-certified, Minimally Trained Respiratory Research Assistants. Curves displaying an "equipment-related" artifact were not among spirograms rejected by Non-certified, Minimally Trained, Respiratory Research Assistants (refer to Figure 6.1).

One-fifth of tracings were judged non-acceptable by Respiratory Epidemiologists (refer to Table 7.1c). Coinciding with findings from Non-certified, Minimally Trained, Respiratory Research Assistants, 45.0% of rejected spirograms failed to satisfy both start-of-test and end-of-test criteria. The remaining 55.0% were considered to indicate end-of-test failure, only. In general, types of explanations for test failure coincided with those of both previous rater categories. Further, the artifact illustrated in Figure 6.1, recognized by Certified, Respiratory Technicians appeared on 25.0% of spirograms deemed non-acceptable. It was also attributed to a technical malfunction.

Pulmonary Specialists evaluated 24.0% of best test spirograms as non-acceptable (refer to Table 7.1d). Approximately 57.0% of these curves failed on account of an unsatisfactory start and termination. Further, rejection of the remaining 43.0% was ascribable to non-compliance with end-of-test criteria. Consistent types of explanations were also recorded. Again, all non-acceptable curves generating an equipment-related comment featured the artifact described in Figure 6.1.

Thus, across expertise categories, the absence of obvious differences in the proportion of non-acceptable spirograms representing start-of-test failure, end-of-test failure, or both indicated that raters with varying levels of respiratory expertise were similarly selective in their application of ATS criteria. However, during the analysis, observation of an equipment related artifact, identified only among spirograms deemed non-acceptable by raters from professional expertise categories, suggested an association between professional, respiratory expertise and the potential to detect patterns indicative of a technical complication.

b) Effects of Technician Expertise on Non-acceptability

Regardless of rater category, approximately equal proportions of tracings from minimally trained and highly trained technicians received "non-acceptable" ratings (p-values of 0.56, 0.67, 0.86, and 1.00 corresponded to the Non-certified, Minimally Trained, Respiratory Research Assistants; Certified, Respiratory Technicians; Respiratory Epidemiologists; and Pulmonary Specialists, respectively). Both the variety and distribution of comments relating to each technician's spirograms were also similar. Consequently, no evidence of a technician effect on raters' explanations of spirogram non-acceptability was detected.

ii) The Secondary Data Set

a) Overall Non-acceptability

Based on the interpretations of Non-certified, Minimally Trained, Respiratory Research Assistants, 20.0% of best test spirograms were not acceptable (refer to Table 7.2a). Slightly over one-half of such tracings (55.0%) displayed features considered to reflect end-of-test failure. Among those remaining, the proportion rejected as a result of both an unsatisfactory test initiation and termination was twice as high as that attributed to start-of-test failure, only. Regarding explanations for non-acceptability, "variable effort", indistinguishable from "coughing", characterized anomalies appearing near the start of curves. "Incomplete exhalation" was the most common explanation of non-acceptable end-of-test patterns.

Certified Respiratory Technicians deemed 28.9% of spirograms non-acceptable (refer to Table 7.2b). Rejection of 35.1%, a proportion similar to that observed by Non-certified, Minimally Trained, Respiratory Research Assistants, was attributed to both start-of-test and end-of-test failure. The majority of remaining tracings (54.4%) exhibited characteristics of an unsatisfactory start. Explanations for one-half of start-of-test failures included “variable effort” and “questionable peak flow”. The remaining half presented patterns defined as a “hesitating start”. Importantly, spirograms exhibiting this latter statement were not among tracings rejected by Non-certified, Minimally Trained, Respiratory Technicians. With respect to test termination, “incomplete exhalation” comprised the primary reason for end-of-test failure.

Approximately 13% of spirograms received a “non-acceptable” rating by both Respiratory Epidemiologists (refer to Table 7.2c). Values of 24.0%, 28.0%, and 48.0% corresponded to proportions of rejected tracings classified as start-of-test failures, end-of-test failures, or both, respectively. Regarding the types of explanations for spirogram rejection, “hesitating start”, “coughing” and “variable effort” defined anomalies exhibited at the beginning of curves and “incomplete exhalation” characterized patterns appearing within the latter portion. Spirograms displaying “hesitating starts” were the same tracings that generated this interpretation by Certified, Respiratory Technicians.

Pulmonary Specialists judged 20% of best test spirograms as not acceptable. These curves were distributed equally across the three types of test failure (i.e., start-of-test, end-of-test, or both) (refer to Table 7.2d). “Variable effort”, “cough”, and “questionable peak flow” comprised the primary explanations for start-of-test failure. Similar to the interpretations of Non-certified, Minimally Trained, Respiratory Research Assistants, no curve pattern was identified as a “hesitating start”. Further, the spirograms classified as exhibiting a “hesitating start” by the previous two expertise levels were not among those considered non-acceptable by the Pulmonary Specialists. Comments of “incomplete exhalation”, “cough”, and “variable effort” (pertaining to end-of-test failure) coincided with statements contributed by raters from all previous categories.

Analogous to trends in the Primary Data Set, similar proportions of non-acceptable spirograms within each category of raters failed to satisfy both ATS start-of-test and end-of-test criteria. Therefore, a comparable selective approach to assigning explanations for test failure was applied by all raters.

b) Effects of Technician Expertise on Non-acceptability

Raters' interpretations were also assessed according to technician expertise. Resembling results of the Primary Data Set, equivalent proportions of spirograms from minimally trained and highly trained technicians were evaluated as non-acceptable by Non-certified, Minimally Trained, Respiratory Research Assistants (p-value = 1.00) (refer to Table 7.2a). The lack of variation in the distribution of "reasons for rejection" across technician expertise levels further suggested the absence of a technician effect on raters' interpretations of test failure. In contrast, across all categories of raters with professional, respiratory expertise, the proportion of rejected spirograms was lower for the highly trained technician, providing evidence of a technician effect (refer to Tables 7.2b, 7.2c, and 7.2d). Forty-one percent and 16.0.% of spirograms from the minimally trained technician and highly trained technician, respectively, were deemed non-acceptable by the Certified Respiratory Technicians (p-value = 0.006). Similarly, the percentage of tracings from minimally trained technicians that received "non-acceptable" ratings by Respiratory Epidemiologists and Pulmonary Specialists was three-fold higher than that calculated for the highly trained technician's spirograms (p-values $\leq$ 0.006). Across all professional rater categories, a significant majority of spirograms failing to strictly meet ATS start-of-test criteria were produced by minimally trained technicians. These trends indicated the existence of a technician effect on the interpretations of acceptability for raters with professional, respiratory expertise.

7.2.2 Variations in Explanations of a Common Spirogram According to Rater Expertise

To determine whether raters' interpretations of a common curve artifact varied according to their respiratory expertise, comments extracted from spirograms judged non-acceptable across all expertise categories were compared.

i) The Primary Data Set

Of the100 best test spirograms, three (one from the minimally trained technician and two from the highly trained technician) were evaluated as non-acceptable by all raters. Curves, coupled with their respective interpretations, are presented in Figures 7.1a, 7.1b, and 7.1c. In each figure, explanations provided by raters within expertise categories referred to a common test point (start-of-test, end-of-test, or both). Regarding Figure 7.1a, statements from all Non-certified, Minimally Trained, Respiratory Research Assistants and all Certified, Respiratory Technicians attributed spirogram non-acceptability to both start-of-test and end-of-test failure. Comments indicative of end-of-test failure corresponded to Respiratory Epidemiologists and Pulmonary Specialists. Consistent trends were noted in Figures 7.1b and 7.1c. However, interpretations varied between expertise categories. Non-certified, Minimally Trained, Respiratory Research Assistants and Certified, Respiratory Technicians deemed the spirogram in Figure 7.1a non-acceptable upon identification of an unsatisfactory start and termination while rejection of this tracing by both Respiratory Epidemiologists and Pulmonary Specialists was completely ascribed to end-of-test failure. Comparably, according to interpretations from Non-certified, Minimally Trained, Respiratory, Research Assistants, Certified, Respiratory Technicians, and Respiratory Epidemiologists, the spirogram illustrated in Figure 7.1b reflected both start-of-test and end-of-test failure, whereas those contributed by Pulmonary Specialists pertained to end-of-test failure, only. Importantly, no two categories, despite agreement on the point of test failure, defined patterns exhibited in any of the three figures similarly. For example, the artifact appearing within the final portion of the tracing in Figure 7.1a, was, depending on expertise category, considered to represent "poor termination", "a leak", "variable effort", or "inhalation". These results suggested that raters' interpretations of spirogram non-acceptability was respiratory expertise-specific.

ii) The Secondary Data Set

Of the 200 best test spirograms, five (four from minimally trained technicians and one from the highly trained technician) were judged non-acceptable by all raters. Tracings and respective interpretations are displayed in Figures 7.2a through 7.2e. In contrast to the results of the Primary Data Set, raters within and across all categories identically defined test failure on each spirogram. In Figures 7.2b, 7.2d, and 7.2e, all statements referred to an unsatisfactory start. Those coinciding with Figures 7.2a and 7.2c attributed spirogram rejection to non-compliance with both start-of-test and end-of-test criteria. Further, identical combinations of explanations for specific curve patterns were recognized within each category. "Cough", "variable effort", and "flow fluctuations" appeared together indicating their interchangeable use among all raters from all expertise levels. Thus, based on these trends, no variance in raters' interpretations of a common spirogram was evident.

7.3 Conclusions

Across the Primary and Secondary Data Sets, analysis of explanations for spirogram non-acceptability indicated that raters, within their individual expertise categories, classified comparable proportions of rejected spirograms as start-of-test failures, end-of-test failures, or both. These results suggested that raters defined spirogram non-acceptability with a similar selective precision regardless of their respiratory expertise. Assessment of the effect of technician expertise on spirometric test failure generated different findings for each data set. Results of the Primary Data Set exhibited no evidence of such an effect. In contrast, those of the Secondary Data Set displayed trends reflecting the presence of a technician effect on the judgements of raters with professional, respiratory expertise. Variations in raters' interpretations of a common spirogram were also inconsistent across data sets. Only results of the Primary Data Set reflected expertise-related differences in raters' explanations for spirogram rejection. Thus, noted discrepancies within and between data sets precluded the establishment of a clear relationship between raters' respiratory expertise and the interpretation of spirometric test failure.

Table 7.1a Reasons Provided by Non-certified, Minimally Trained, Respiratory Research Assistants for Evaluating Best Test Spirograms as Non-acceptable in the Primary Data Set

	Expertise Level of Technician Conducting Spirometric Test		
Reasons For Rejecting Best Test Spirogram	Minimally Trained Technician	Highly Trained Technician	
Failure to Meet ATS Start of Test Criteria:			
Inhalation at Start	-----	-----	
Hesitating Start	-----	-----	
Variable Effort / Slow Start	-----	-----	
Variable Effort / Cough	------	-----	
Variable Effort / Questionable Peak Flow	-----	-----	
Cough / No Peak Flow	-----	-----	
Poor Start of Test	-----	-----	
Failure to Meet ATS End of Test Criteria:			
Poor Termination	1	1	
Poor Termination - Possible Technical Problem	-----	-----	
Incomplete Exhalation	1	3	
Inhalation	-----	1	
Leak	-----	-----	
Variable Effort	1	-----	
Failure to Meet Both ATS Start of Test and End of Test Criteria:			
Poor Start of Test and Questionable Exhalation/Termination	1	1	
Variable Effort at Start of Test and Incomplete Exhalation/Termination	-----	-----	
Possible Cough / Variable Effort at Start of Test and Incomplete Exhalation / Termination	1	1	
Hesitating start and incomplete exhalation	1	1	
Poor start and leak	-----	-----	
Flow Fluctuations and Variable Effort Throughout Test	-----	1	
Cough or Flow Fluctuations Throughout Test	-----	-----	
Technical Problem / Inhalation	-----	-----	
Number of Best Test Spirograms Evaluated as Non-acceptable	6 (12.8)[†]	9 (17.0)[†]	p-value* = 0.556

* p-value based on chi-square test statistic

[†] Percent based on total number of best test spirograms (n = 100)

Note: Statements provided by different raters regarding the non-acceptability of a common spirogram are separated by a "/".

Table 7.1b Reasons Provided by Certified Respiratory Technicians for Evaluating Best Test Spirograms as Non-acceptable in the Primary Data Set

	Expertise Level of Technician Conducting Spirometric Test		
Reasons For Rejecting Best Test Spirogram	Minimally Trained Technician	Highly Trained Technician	
Failure to Meet ATS Start of Test Criteria:			
Inhalation at Start	-----	-----	
Hesitating Start	-----	-----	
Variable Effort / Slow Start	-----	1	
Variable Effort / Cough	2	1	
Variable Effort / Questionable Peak Flow	-----	-----	
Cough / No Peak Flow	-----	-----	
Poor Start of Test	-----	-----	
Failure to Meet ATS End of Test Criteria:			
Poor Termination	-----	-----	
Poor Termination - Possible Technical Problem	1	2	
Incomplete Exhalation	-----	1	
Inhalation	1	1	
Leak	-----	-----	
Variable Effort	-----	-----	
Failure to Meet Both ATS Start of Test and End of Test Criteria:			
Poor Start of Test and Questionable Exhalation/Termination	-----	-----	
Variable Effort at Start of Test and Incomplete Exhalation/Termination	1	1	
Possible Cough or Variable Effort at Start of Test and Incomplete Exhalation / Termination	1	-----	
Hesitating start and incomplete exhalation	-----	-----	
Poor start and leak	1	-----	
Flow Fluctuations and Variable Effort Throughout Test	1	3	
Coughing or Flow Fluctuations Throughout Test	1	2	
Technical Problem / Inhalation	-----	-----	
Number of Best Test Spirograms Evaluated as Non-acceptable	9 (19.1)[†]	12 (22.6)[†]	p-value* = 0.669

* p-value based on chi-square test statistic

[†] Percent based on total number of best test spirograms (n = 100)

Note: Statements provided by different raters regarding the non-acceptability of a common spirogram are separated by a "/".

Table 7.1c Reasons Provided by Respiratory Epidemiologists for Evaluating Best Test Spirograms as Non-acceptable in the Primary Data Set

	Expertise Level of Technician Conducting Spirometric Test		
Reasons For Rejecting Best Test Spirogram	Minimally Trained Technician	Highly Trained Technician	
Failure to Meet ATS Start of Test Criteria:			
Inhalation at Start	-----	-----	
Hesitating Start	-----	-----	
Variable Effort / Slow Start	-----	-----	
Variable Effort / Cough	-----	-----	
Variable Effort / Questionable Peak Flow	-----	-----	
Cough / No Peak Flow	-----	-----	
Poor Start of Test	-----	-----	
Failure to Meet ATS End of Test Criteria:			
Poor Termination / Questionable Termination	-----	1	
Poor Termination - Possible Technical Problem	2	2	
Incomplete Exhalation	1	2	
Inhalation	1	1	
Leak	-----	-----	
Variable Effort	1	-----	
Failure to Meet Both ATS Start of Test and End of Test Criteria:			
Poor Start of Test and Questionable Exhalation/Termination	1	1	
Variable Effort at Start of Test and Incomplete Exhalation/Termination	-----	-----	
Possible Cough or Variable Effort at Start of Test and Incomplete Exhalation / Termination	1	-----	
Hesitating start and incomplete exhalation	-----	-----	
Poor start and leak.	-----	-----	
Flow Fluctuations and Variable Effort Throughout Test	2	4	
Cough or Flow Fluctuations Throughout Test	-----	-----	
Technical Problem / Inhalation	-----	-----	
Number of Best Test Spirograms Evaluated as Non-acceptable	9 (19.1)[†]	11 (20.8)[†]	p-value[*] = 0.857

[*] p-value based on chi-square test statistic

[†] Percent based on total number of best test spirograms (n = 100)

Note: Statements provided by different raters regarding the non-acceptability of a common spirogram are separated by a "/".

Table 7.1d Reasons Provided by Pulmonary Specialists for Evaluating Best Test Spirograms as Non-acceptable in the Primary Data Set

	Expertise Level of Technician Conducting Spirometric Test		
Reasons For Rejecting Best Test Spirogram	Minimally Trained Technician	Highly Trained Technician	
Failure to Meet ATS Start of Test Criteria:			
Inhalation at Start	-----	-----	
Hesitating Start	-----	-----	
Variable Effort / Slow Start	-----	-----	
Variable Effort / Cough	-----	-----	
Variable Effort / Questionable Peak Flow	-----	-----	
Cough / No Peak Flow	-----	-----	
Poor Start of Test	-----	-----	
Failure to Meet ATS End of Test Criteria:			
Poor Termination / Questionable Termination	-----	-----	
Poor Termination - Possible Technical Problem	6	3	
Incomplete Exhalation	3	3	
Inhalation	-----	1	
Leak	-----	-----	
Variable Effort	-----	-----	
Failure to Meet Both ATS Start of Test and End of Test Criteria:			
Poor Start of Test and Questionable Exhalation/Termination	-----	-----	
Variable Effort at Start of Test and Incomplete Exhalation/Termination	-----	-----	
Possible Cough or Variable Effort at Start of Test and Incomplete Exhalation / Termination	1	1	
Hesitating start and incomplete exhalation	-----	-----	
Poor start and leak	-----	-----	
Flow Fluctuations and Variable Effort Throughout Test	3	1	
Coughing or Flow Fluctuations Throughout Test	-----	-----	
Technical Problem / Inhalation	1	1	
Number of Best Test Spirograms Evaluated as Non-acceptable	14 (30.0)†	10 (18.9)†	p-value* = 0.202

* p-value based on chi-square test statistic

† Percent based on total number of best test spirograms (n = 100)

Note: Statements provided by different raters regarding the non-acceptability of a common spirogram are separated by a "/".

Table 7.2a Reasons Provided by Non-certified, Minimally Trained, Respiratory Research Assistants for Evaluating Best Test Spirograms as Non-acceptable in the Secondary Data Set

	Expertise Level of Technician Conducting Spirometric Test		
Reasons For Rejecting Best Test Spirogram	Minimally Trained Technician	Highly Trained Technician	
Failure to Meet ATS Start of Test Criteria:			
Inhalation at Start	-----	-----	
Hesitating Start	-----	-----	
Variable Effort / Slow Start	-----	-----	
Variable Effort / Cough	3	3	
Variable Effort / Questionable Peak Flow	-----	-----	
Cough / No Peak Flow	-----	-----	
Poor Start of Test	-----	-----	
Failure to Meet ATS End of Test Criteria:			
Poor Termination / Questionable Termination	2	2	
Poor Termination - Possible Technical Problem	-----	-----	
Incomplete Exhalation	8	8	
Inhalation	-----	1	
Leak	-----	-----	
Variable Effort	1	-----	
Failure to Meet Both ATS Start of Test and End of Test Criteria:			
Poor Start of Test and Questionable Exhalation/Termination	-----	-----	
Variable Effort at Start of Test and Incomplete Exhalation/Termination	1	2	
Possible Cough or Variable Effort at Start of Test and Incomplete Exhalation / Termination	3	3	
Hesitating Start and incomplete exhalation	-----	-----	
Poor Start and Leak	-----	-----	
Flow Fluctuations and Variable Effort Throughout Test	2	1	
Coughing or Flow Fluctuations Throughout Test	-----	-----	
Technical Problem / Inhalation	-----	-----	
Number of Best Test Spirograms Evaluated as Nonacceptable	20 (20.0)†	20 (20.0)†	p-value* = 1.000

* p-value based on chi-square test statistic
† Percent based on total number of best test spirograms (n = 200)

Note: Statements provided by different raters regarding the non-acceptability of a common spirogram are separated by a "/".

Table 7.2b Reasons Provided by Certified Respiratory Technicians for Evaluating Best Test Spirograms as Non-acceptable in the Secondary Data Set

	Expertise Level of Technician Conducting Spirometric Test		
Reasons For Rejecting Best Test Spirogram	Minimally Trained Technician	Highly Trained Technician	
Failure to Meet ATS Start of Test Criteria:			
Inhalation at Start	-----	-----	
Hesitating Start	11	-----	
Variable Effort / Slow Start	-----	-----	
Variable Effort / Cough	-----	1	
Variable Effort / Questionable Peak Flow	12	3	
Cough / No Peak Flow	-----	-----	
Poor Start of Test	4	-----	
Failure to Meet ATS End of Test Criteria:			
Poor Termination / Questionable Termination	-----	-----	
Poor Termination - Possible Technical Problem	-----	-----	
Incomplete Exhalation	3	3	
Inhalation	-----	-----	
Leak	-----	-----	
Variable Effort	-----	-----	
Failure to Meet Both ATS Start of Test and End of Test Criteria:			
Poor Start of Test and Questionable Exhalation/Termination	-----	-----	
Variable Effort at Start of Test and Incomplete Exhalation/Termination	2	1	
Possible Cough or Variable Effort at Start of Test and Incomplete Exhalation / Termination	1	1	
Hesitating Start and incomplete exhalation	-----	-----	
Poor Start and Leak	-----	-----	
Flow Fluctuations and Variable Effort Throughout Test	8	7	
Coughing or Flow Fluctuations Throughout Test	-----	-----	
Technical Problem / Inhalation	-----	-----	
Number of Best Test Spirograms Evaluated as Non-acceptable	41 (41.0)'	16 (16.0)'	p-value* = 0.006

* p-value based on chi-square test statistic

' Percent based on total number of best test spirograms (n = 200)

Note: Statements provided by different raters regarding the non-acceptability of a common spirogram are separated by a "/".

Table 7.2c Reasons Provided by Respiratory Epidemiologists for Evaluating Best Test Spirograms as Non-acceptable in the SecondaryData Set

	Expertise Level of Technician Conducting Spirometric Test		
Reasons For Rejecting Best Test Spirogram	Minimally Trained Technician	Highly Trained Technician	
Failure to Meet ATS Start of Test Criteria:			
Inhalation at Start	-----	-----	
Hesitating Start	1	-----	
Variable Effort / Slow Start	2	-----	
Variable Effort / Cough	3	-----	
Variable Effort / Questionable Peak Flow	-----	-----	
Cough / No Peak Flow	1	1	
Poor Start of Test	-----	-----	
Failure to Meet ATS End of Test Criteria:			
Poor Termination / Questionable Termination	2	1	
Poor Termination - Possible Technical Problem	-----	-----	
Incomplete Exhalation	3	2	
Inhalation	-----	-----	
Leak	-----	-----	
Variable Effort	-----	-----	
Failure to Meet Both ATS Start of Test and End of Test Criteria:			
Poor Start of Test and Questionable Exhalation/Termination	-----	-----	
Variable Effort at Start of Test and Incomplete Exhalation/Termination	1	1	
Possible Cough or Variable Effort at Start of Test and Incomplete Exhalation / Termination	2	1	
Hesitating Start and incomplete exhalation	-----	-----	
Poor Start and Leak	-----	-----	
Flow Fluctuations and Variable Effort Throughout Test	3	-----	
Coughing or Flow Fluctuations Throughout Test	1	-----	
Possible Technical Problem / Inhalation	-----	-----	
Number of Best Test Spirograms Evaluated as Non-acceptable	19 (19.0)[†]	6 (6.0)[†]	p-value* = 0.006

* p-value based on chi-square test statistic

[†] Percent based on total number of best test spirograms (n = 200)

Note: Statements provided by different raters regarding the non-acceptability of a common spirogram are separated by a "/".

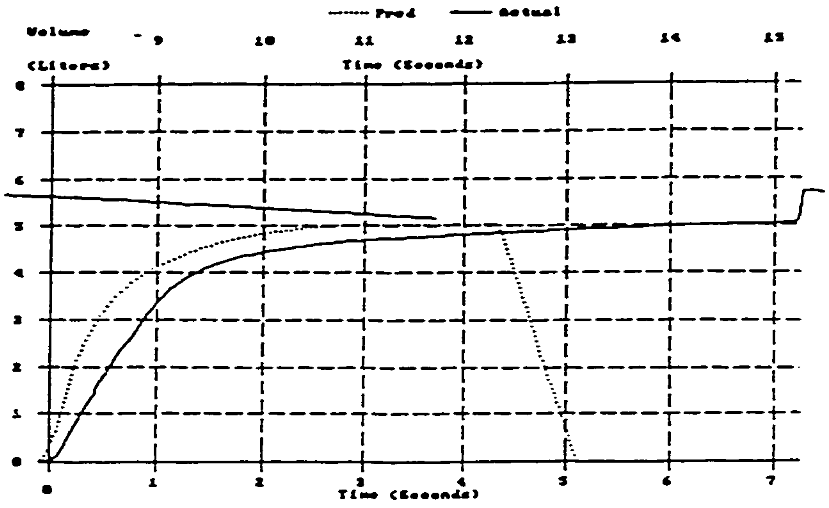
Table 7.2d Reasons Provided by Pulmonary Specialists for Evaluating Best Test Spirograms as Non-acceptable in the Secondary Data Set

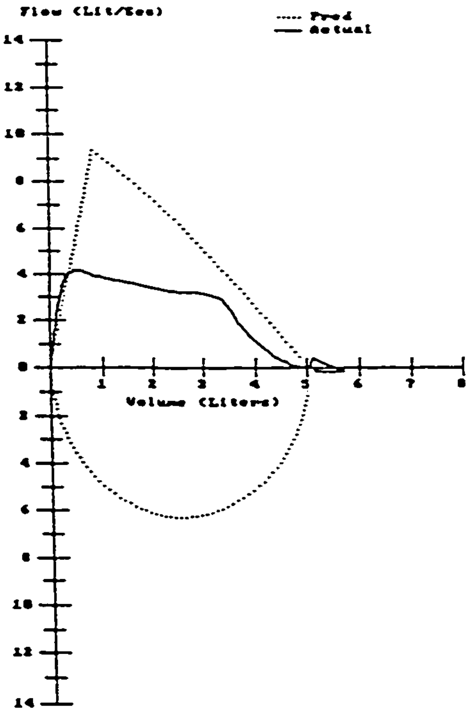
	Expertise Level of Technician Conducting Spirometric Test		
Reasons For Rejecting Best Test Spirogram	Minimally Trained Technician	Highly Trained Technician	
Failure to Meet ATS Start of Test Criteria:			
Inhalation at Start	-----	-----	
Hesitating Start	-----	-----	
Variable Effort / Slow Start	4	-----	
Variable Effort / Cough	3	-----	
Variable Effort / Questionable Peak Flow	4	-----	
Cough / No Peak Flow	2	1	
Poor Start of Test	2	-----	
Failure to Meet ATS End of Test Criteria:			
Poor Termination / Questionable Termination	4	2	
Poor Termination - Possible Technical Problem	-----	-----	
Incomplete Exhalation	2	3	
Inhalation	2	1	
Leak	-----	-----	
Variable Effort	-----	-----	
Failure to Meet Both ATS Start of Test and End of Test Criteria:			
Poor Start of Test and Questionable Exhalation/Termination	-----	-----	
Variable Effort at Start of Test and Incomplete Exhalation/Termination	1	-----	
Possible Cough or Variable Effort at Start of Test and Incomplete Exhalation / Termination	-----	-----	
Poor Start and Leak	-----	-----	
Flow Fluctuations and Variable Effort Throughout Test .	-----	-----	
Flow Fluctuations or Variable Effort Throughout Test	6	2	
Coughing or Flow Fluctuations Throughout Test	1	-----	
Technical Problem / Inhalation	-----	-----	
Number of Best Test Spirograms Evaluated as Non-acceptable	31 (31.0)[†]	9 (9.0)[†]	p-value* = 0.000

* p-value based on chi-square test statistic

[†] Percent based on total number of best test spirograms (n = 200)

Note: Statements provided by different raters regarding the non-acceptability of a common spirogram are separated by a "/".





Technician expertise: Minimally Trained

Interpretations from each rater:
(based upon codes presented in Appendix 15)

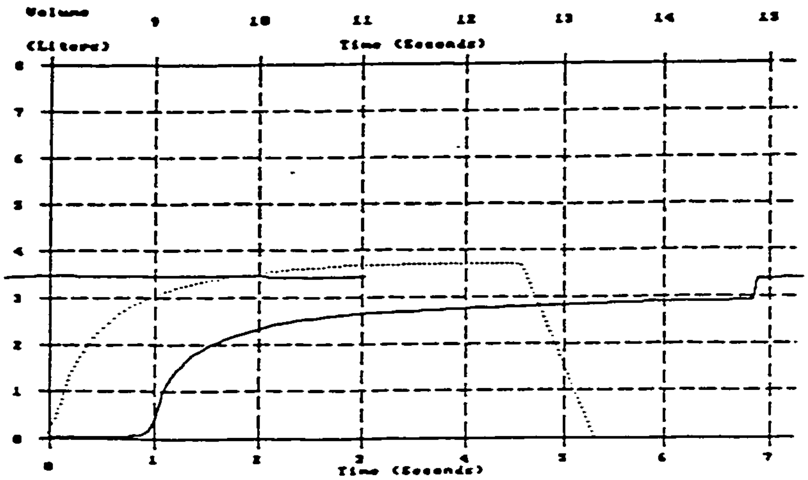
Non-certified, Minimally Trained, Respiratory Research Assistants:
Rater 1: Poor start and poor termination
Rater 2: Poor start and poor termination
Rater 3: Poor start and questionable termination

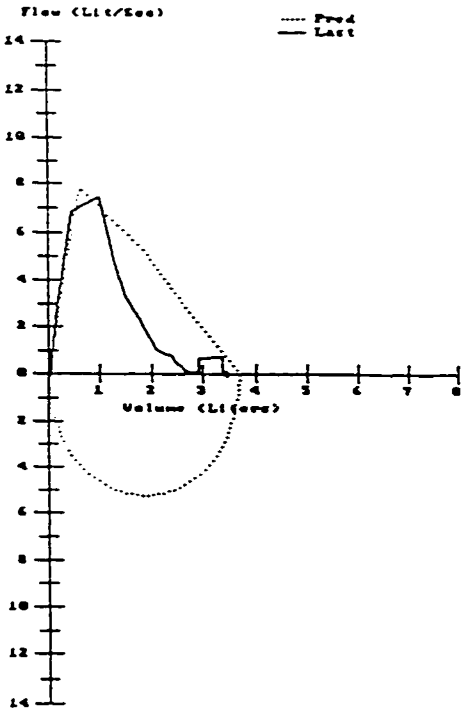
Certified, Respiratory Technicians:
Rater 4: Poor start and leak
Rater 5: Poor start and questionable termination
Rater 6: Questionable start of test and leak

Respiratory Epidemiologists:
Rater 7: Poor termination
Rater 8: Variable effort at end of test

Pulmonary Specialists:
Rater 9: Inhalation
Rater 10: Inhalation

Figure 7.1a Non-acceptable spirogram (1p) and corresponding interpretations from the Primary Data Set





Technician expertise: Highly Trained

Interpretations from each rater:
(based upon codes presented in Appendix 15)

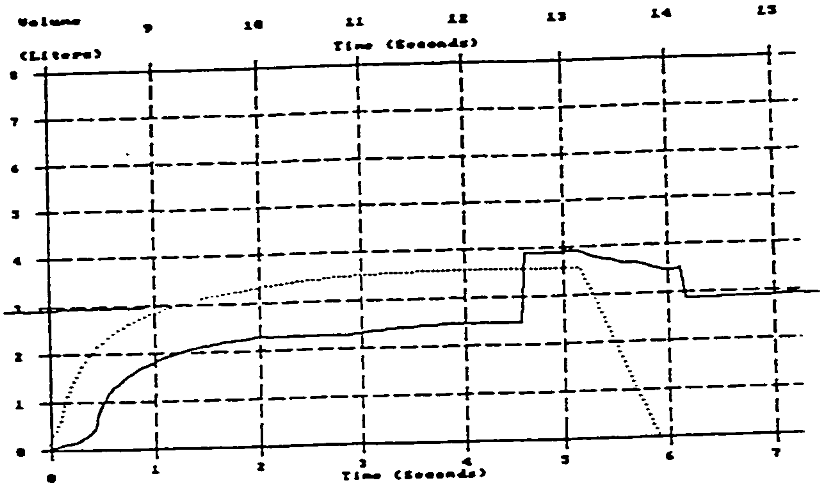
Non-certified, Minimally Trained, Respiratory Research Assistants:
Rater 1: Hesitating start and poor termination
Rater 2: Hesitating start and early termination
Rater 3: Hesitating start and early termination

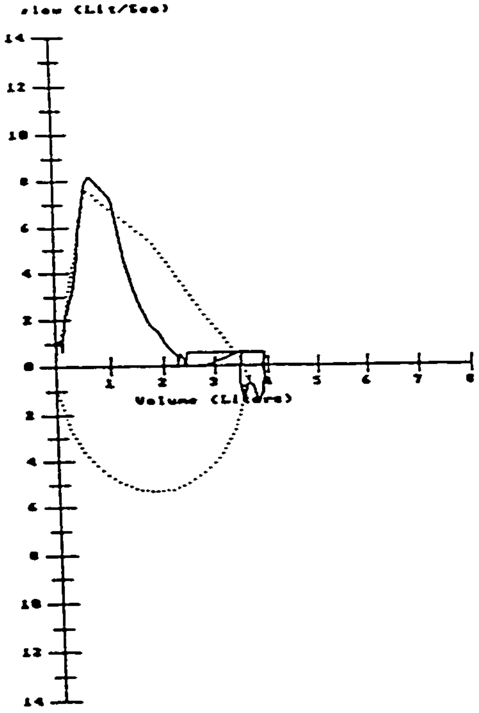
Certified, Respiratory Technicians:
Rater 4: Variable effort at start
Rater 5: Slow start
Rater 6: Variable effort at start

Respiratory Epidemiologists:
Rater 7: Incomplete exhalation
Rater 8: Incomplete exhalation

Pulmonary Specialists:
Rater 9: Inhalation
Rater 10: Technical Problem

Figure 7.1b Non-acceptable spirogram (2p) and corresponding interpretations from the Primary Data Set





Technician expertise: Highly Trained

Interpretations from each rater:
(based upon codes presented in Appendix 15)

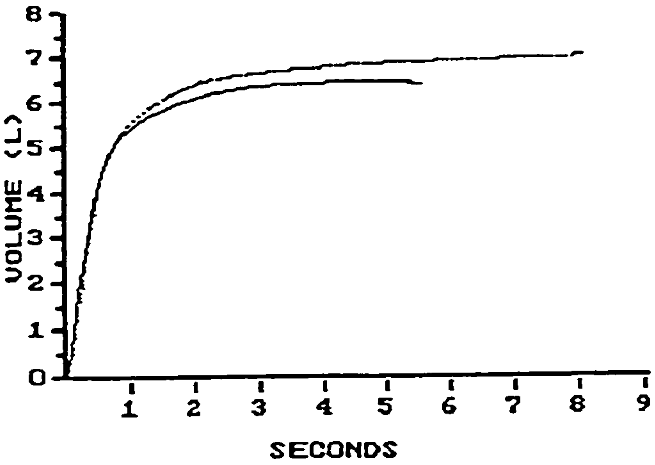
Non-certified, Minimally Trained, Respiratory Research Assistants:
Rater 1: Poor start and poor termination
Rater 2: Poor start and poor exhalation
Rater 3: Questionable start and poor termination

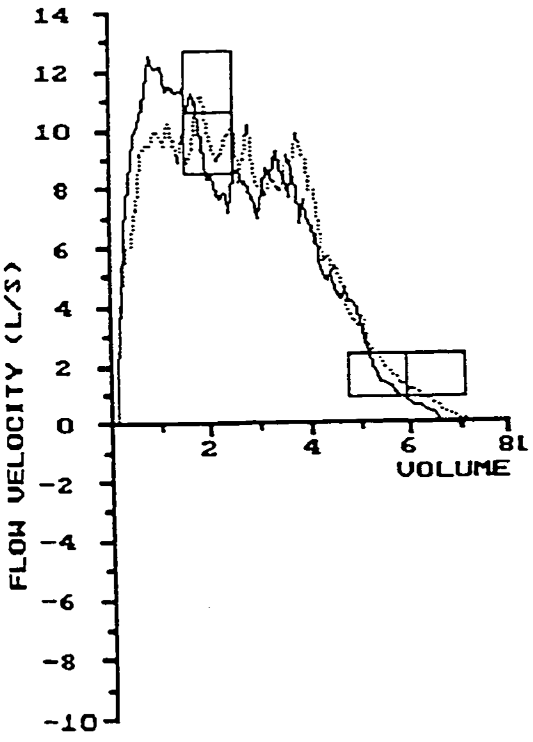
Certified, Respiratory Technicians:
Rater 4: Variable effort throughout
Rater 5: Variable effort and incomplete exhalation
Rater 6: Poor start and questionable termination

Respiratory Epidemiologists:
Rater 7: Cough at start and insufficient exhalation
Rater 8: Variable start and incomplete termination

Pulmonary Specialists:
Rater 9: Poor termination
Rater 10: Technical Problem

Figure 7.1c Non-acceptable spirogram (3p) and corresponding interpretations from the Primary Data Set





Technician expertise: Minimally Trained

Interpretations from each rater:
(based upon codes presented in Appendix 15)

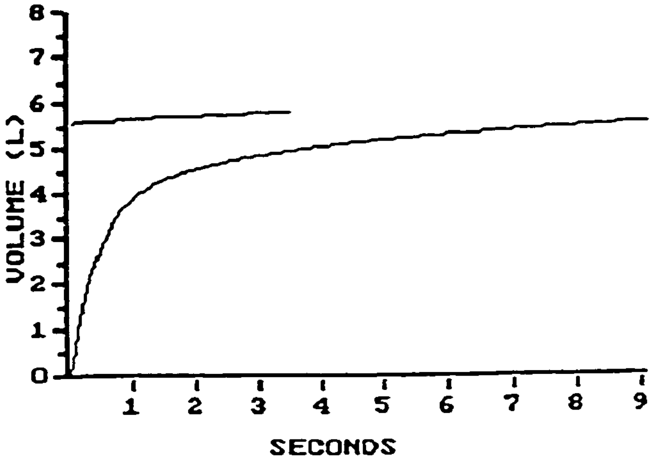
Non-certified, Minimally Trained, Respiratory Research Assistants:
Rater 1: Cough at start and incomplete exhalation
Rater 2: Variable effort at start and poor exhalation
Rater 3: Variable effort at start and poor exhalation

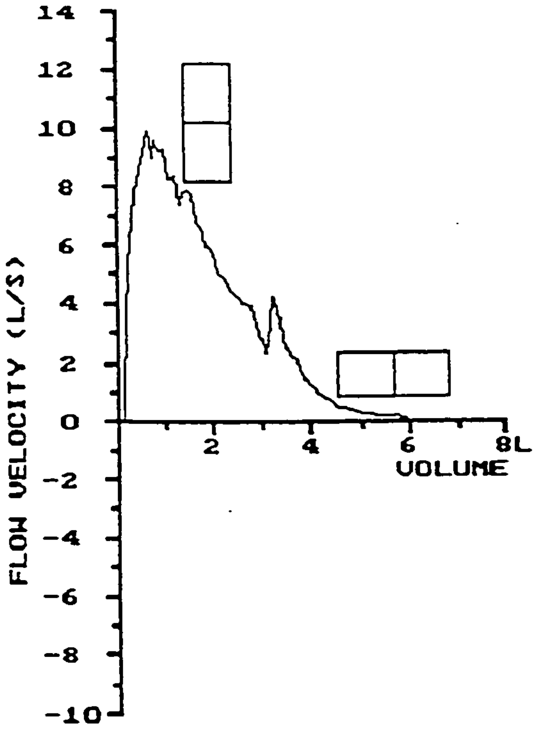
Certified, Respiratory Technicians:
Rater 4: Cough at start and incomplete exhalation
Rater 5: Cough at start and poor termination
Rater 6: Variable effort at start and incomplete exhalation

Respiratory Epidemiologists:
Rater 7: Coughing throughout test
Rater 8: Flow fluctuations throughout test

Pulmonary Specialists:
Rater 9: Flow fluctuations throughout test
Rater 10: Coughing throughout test

Figure 7.2a Non-acceptable spirogram (1s) and corresponding interpretations from the Secondary Data Set





Technician expertise: Minimally Trained

Interpretations from each rater:
(based upon codes presented in Appendix 15)

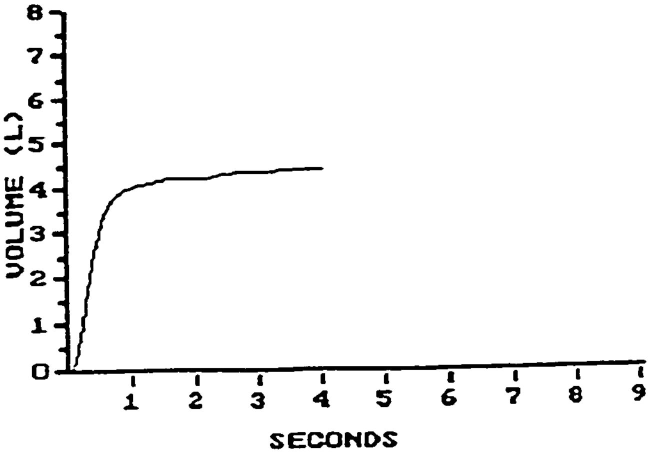
Non-certified, Minimally Trained, Respiratory Research Assistants:
Rater 1: Cough at start
Rater 2: Cough at start
Rater 3: Cough at start

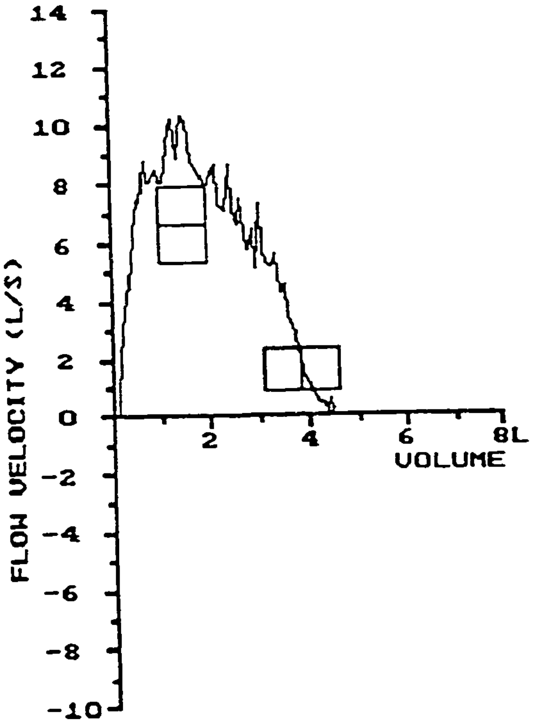
Certified, Respiratory Technicians:
Rater 4: Variable effort at start
Rater 5: Variable effort at start
Rater 6: Variable effort at start

Respiratory Epidemiologists:
Rater 7: Cough at start
Rater 8: Cough at start

Pulmonary Specialists:
Rater 9: Cough at start
Rater 10: Cough at start

Figure 7.2b Non-acceptable spirogram(2s) and corresponding interpretations from the Secondary Data Set





Technician expertise: Minimally Trained

Interpretations from each rater:
(based upon codes presented in Appendix 15)

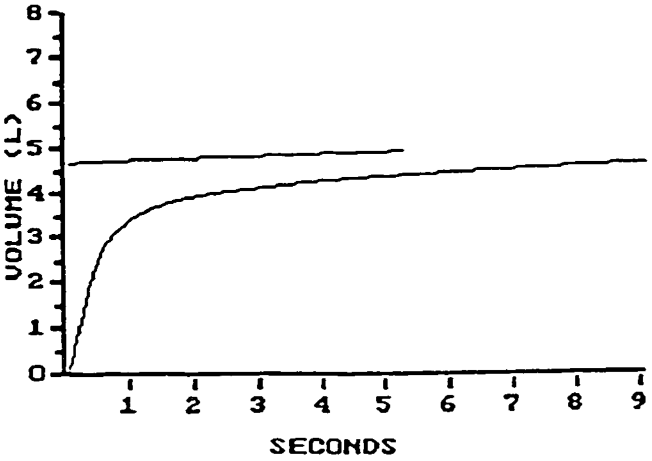
Non-certified, Minimally Trained, Respiratory Research Assistants:
Rater 1: Poor start and incomplete termination
Rater 2: Variable effort at start and incomplete termination
Rater 3: Cough at start and incomplete termination

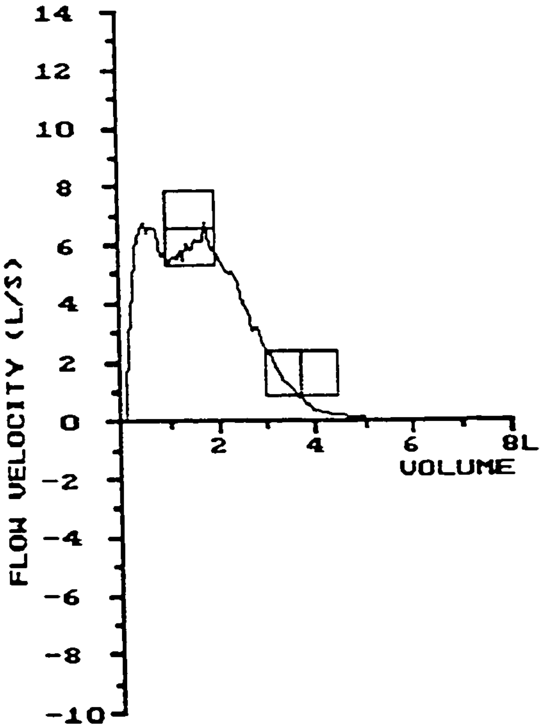
Certified, Respiratory Technicians:
Rater 4: Variable effort at start and incomplete exhalation
Rater 5: Variable effort and incomplete exhalation
Rater 6: Poor start and questionable termination

Respiratory Epidemiologists:
Rater 7: Flow fluctuations throughout test
Rater 8: Variable effort throughout test

Pulmonary Specialists:
Rater 9: Variable effort at start and incomplete termination
Rater 10: Variable throughout test

Figure 7.2c Non-acceptable spirogram (3s) and corresponding interpretations from the Secondary Data Set





Technician expertise: Minimally Trained

Interpretations from each rater:
(based upon codes presented in Appendix 15)

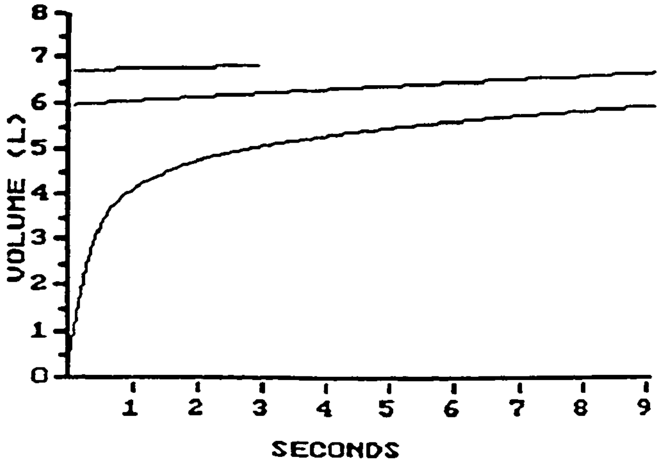
Non-certified, Minimally Trained, Respiratory Research Assistants:
Rater 1: Variable effort at start
Rater 2: Variable effort at start
Rater 3: Cough at start

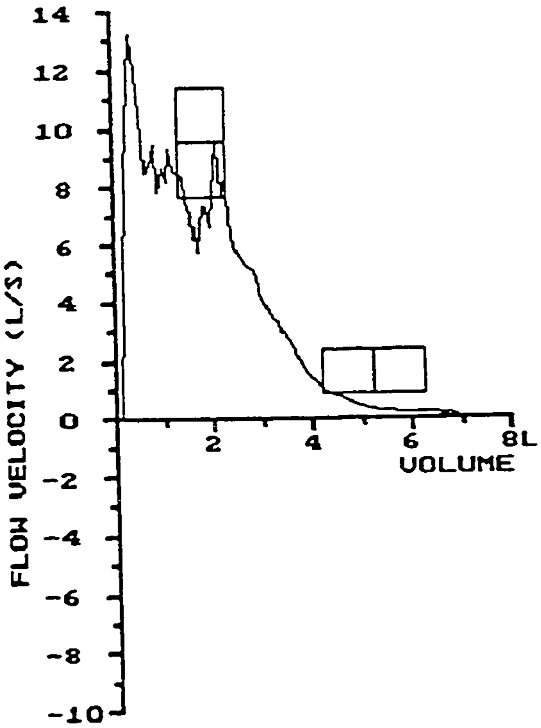
Certified, Respiratory Technicians:
Rater 4: Variable effort at start
Rater 5: Poor start
Rater 6: Variable effort at start

Respiratory Epidemiologists:
Rater 7: Cough at start
Rater 8: Variable effort at start

Pulmonary Specialists:
Rater 9: Questionable peak flow
Rater 10: Variable effort at start

Figure 7.2d Non-acceptable spirogram (4s) and corresponding interpretations from the Secondary Data Set





Technician expertise: Highly Trained

Interpretations from each rater:
(based upon codes presented in Appendix 15)

Non-certified, Minimally Trained, Respiratory Research Assistants:
Rater 1: Poor start
Rater 2: Variable effort at start
Rater 3: Cough at start

Certified, Respiratory Technicians:
Rater 4: Variable effort at start
Rater 5: Cough at start
Rater 6: Variable effort at start

Respiratory Epidemiologists:
Rater 7: Cough at start
Rater 8: Cough at start

Pulmonary Specialists:
Rater 9: Cough at start
Rater 10: Cough at start

Figure 7.2e Non-acceptable spirogram (5s) and corresponding interpretations from the Secondary Data Set

CHAPTER EIGHT

RESULTS - RELATIONSHIPS BETWEEN PARTICIPANT CHARACTERISTICS AND SPIROMETRIC TEST ACCEPTABILITY

In contrast to previous chapters which assessed the influence of raters' respiratory expertise on the evaluation of spirometric results, this chapter examines associations between characteristics of participants performing spirometric maneuvers and test acceptability.

8.1 Definition of Spirogram Acceptability

Bivariate and multivariate analyses of participant characteristics (i.e., factors) potentially related to the achievement of acceptable spirometric results (i.e., the outcome) were completed. For these analyses, test acceptability was defined according to the interpretations of Respiratory Epidemiolgists and Pulmonary Specialists. In respiratory epidemiologic studies conducted within clinical or field settings, pulmonary function data are typically reviewed by medical directors of affiliated pulmonary function laboratories (Pulmonary Physiologists or Pulmonary Physicians) or Respiratory Epidemiologists. Therefore, best test spirograms that received "acceptable" ratings across both expertise levels were deemed acceptable.

8.2 Results of Bivariate Analyses

To screen each data set for significant variables prior to applying multivariate methods, a bivariate analysis was performed. The significance of resultant odds ratios[1] were assessed, establishing the presence/absence of relationships between participant variables and spirometric test acceptability.

[1] An odds ratio is defined as the ratio of the odds of exposure (i.e., manifesting a certain characteristic) among cases (participants producing an acceptable spirometric test) to that among controls (participants producing a non-acceptable spirometric test) (Hennekens, 1987).

8.2.1 Demographic Related Factors and Spirometric Test Acceptability

i) The Primary Data Set

Associations between demographic factors and test acceptability are presented in Table 8.1. No difference in test acceptability by gender was observed (OR:1.00; 95 % C.I.: 0.44 - 2.29). Similarly, the average age of participants who produced acceptable spirograms did not differ significantly from that of participants whose tracings were rejected (p-value = 0.114). Detection of a statistically non-significant increased risk of test failure in participants over the age of 49 years further suggested the absence of a relationship between test acceptability and age (The odds ratio for producing an acceptable test among participants over the age of 49 years was 0.49; 95% C.I.: 0.20 - 1.20). Test acceptability was not related to smoking status. All respective 95% confidence intervals included the null value of 1.0, indicating that observed relationships were not statistically significant at the 0.05 level. Therefore, no demographic characteristic of participants appeared to be associated with test acceptability.

ii) The Secondary Data Set

Parallel analysis of participants' demographic characteristics in the Secondary Data Set are summarized in Table 8.2. Neither gender, age, nor smoking status varied with test acceptability (all corresponding 95% confidence intervals spanned the null value).

8.2.2 Cardiopulmonary Health Related Factors and Spirometric Test Acceptability

i) The Primary Data Set

Presentation with a history of shortness of breath, painful breathing, wheeze or asthma, or frequent cough did not prove to be associated with spirometric test failure (once again, all 95% confidence intervals included the null value of 1.0) (refer to Table 8.1). Similarly, no statistically significant relationship between report of an abnormal chest x-ray and test acceptability was noted. Only one participant indicated a history of lung surgery or heart disease. Consequently, risk estimates could not be calculated for either variable.

In general, the number of positive responses for cardiopulmonary variables was too small to produce reliable estimates of their associations with test acceptability.

ii) The Secondary Data Set

As mentioned in Chapter Two, no cardiopulmonary health-related information was collected for participants, preventing comparisons between the two data sets.

8.2.3. Spirometry-Related Factors and Spirometric Test Acceptability

i) The Primary Data Set

Bivariate analyses of factors pertaining to participants' spirometric test sessions are outlined in Table 8.1. Substantiating findings from previous chapters, no relationship between the expertise level of technicians administering spirometry and test acceptability was noted (OR: 1.00; 95 % C.I.: 0.44 - 2.25). Additionally, percent of predicted lung function values less than the lower limit of normal did not appear to be associated with acceptable tests, although the odds ratio was elevated (OR: 2.03; 95 % C.I.: 0.39 - 10.66). However, small sample sizes contributed to wide confidence intervals and low statistical power[1]. No correlation between test acceptability and compliance with ATS reproducibility criteria was detected [i.e., best test spirograms extracted from ATS-reproducible spirometry sessions were no more likely to be deemed acceptable (OR: 1.45; 95 % C.I.: 0.61 - 3.46)]. Similarly, test acceptability did not vary with a participant's past exposure to spirometry (i.e., completion of spirometric tests prior to involvement in the present study), suggesting the absence of a practice or learning effect (OR: 1.88; C.I.: 0.63 - 5.47). In contrast, participants, whose complementary spirograms were acceptable, exhibited a significantly increased probability of producing an acceptable best test spirogram (OR: 33.89; 95 % C.I.: 10.33 - 111.19). Thus, with the exception of complementary spirogram acceptability, no statistically significant relationship between spirometry-related variables examined and test acceptability was established.

[1] Power refers to the ability of a study to detect true differences when they exist (Last, 1995).

ii) The Secondary Data Set

Three spirometry-related factors significantly correlated with test acceptability (refer to Table 8.2). Tests conducted by highly trained technicians demonstrated a greater likelihood of receiving an acceptable rating than those administered by minimally trained technicians (OR: 3.04; 95 % C.I.: 1.71 - 5.41). Increased test acceptability corresponded to best test spirograms from ATS-reproducible test sessions (OR: 2.54; 95 % C.I.:1.37 - 4.73). Resembling the results of the Primary Data Set, production of acceptable complementary spirograms was positively associated with best test acceptability (OR: 6.34; 95 % C.I.: 3.44 - 11.73).

8.2.4 Confounding and Effect Modification

Observed relationships were assessed for confounding and effect modification. Confounding is defined as a distortion in the estimated effect of an exposure (i.e., an independent variable) on an outcome (i.e., the dependent variable) resulting from the existence of an extraneous factor that is both associated with the exposure and, independently, a determinant of the outcome (Hennekens, 1987). Effect modification refers to variation in the magnitude of an exposure effect across levels of another factor (Rothman,1998). Detection of confounding and effect modification often involves stratification techniques whereby the strength of relationships between exposure and outcome are analyzed in well-defined, homogenous categories (strata) of the confounding variable (Hennekens, 1987).

In the present study, relationships between various factors were examined upon stratification of each data set's study population by potential confounders/effect modifiers. Populations were stratified by variables identified either in the previous analysis or in the literature as affecting test acceptability. They included age over 49 years, expertise level of the spirometry technician, acceptability of complementary spirograms, and reproducibility of the best test spirogram's numeric results. These variables were also considered to be associated with each other. For example, a relationship between technician expertise and both complementary spirogram acceptability and test reproducibility was recognized in the results of the Secondary Data Set. Additionally, as discussed in Chapter Three, a higher proportion of participants aged 50 years and older was initially tested by the highly trained technician. Previous studies have suggested an increased risk of spirometric test failure in subjects over 50 years of age (Hankinson, 1991).

For each potential confounding factor, stratum-specific (unconfounded) odds ratios were calculated and compared with overall crude odds ratios and Mantel-Haenszel pooled odds ratios (Mantel and Haenszel, 1959). Small sample sizes in many of the strata may have resulted in the production of unreliable odds ratios. Corresponding overlapping confidence intervals were large and frequently included the null value. Thus, it was difficult to evaluate the similarity of stratum-specific odds ratios. This became important when attempting to classify the association as either confounding or effect modification since the two are differentiated from one another based, respectively, on the presence or absence of homogeneity (i.e., uniformity) in odds ratios across strata (Hennekens, 1987).

i) The Primary Data Set

Findings from stratified analyses of age category, expertise level of the spirometry technician, reproducibility of the best test spirogram's numeric results, and acceptability of the complementary spirogram are presented in Tables 8.3, 8.4, 8.5, and 8.6, respectively. Apart from those pertaining to acceptability of the complementary spirogram, stratum-specific odds ratios remained statistically non-significant (95% confidence intervals spanned the null value). In general, observed differences in the magnitude of values across "exposure levels" might have resulted from small sample sizes within most strata. Although stratification by technician expertise produced patterns suggesting effect modification of the relationship between ATS-reproducibility and test acceptability (i.e., increased test acceptability was associated with ATS-reproducible spirograms generated by highly trained technicians) no reliable conclusions could be formulated (refer to Table 8.4). A similar trend was noted for the effect of technician expertise on test acceptability subsequent to stratification by ATS-reproducibility (refer to Table 8.5). Significant associations between acceptability of the complementary spirogram and best test acceptability existed both prior to and after stratification by all potential risk factors. Since confidence intervals overlapped, stratum-specific odds ratios were regarded as similar. Their non-uniform appearance may be attributable to small sample sizes rather than effect modification.

ii) The Secondary Data Set

Odds ratios, stratified by age over 49 years, expertise level of the spirometry technician, reproducibility of the best test spirogram's numeric results, and acceptability of the complementary spirogram are displayed in Tables 8.7, 8.8, 8.9, and 8.10, respectively. The apparent "diluted effect" of test reproducibility on test acceptability in participants over the age of 49 years was attributed to small cell numbers as opposed to effect modification (refer to Table 8.7). Stratum-specific variations in the degree to which both complementary spirogram acceptability and technician expertise positively influenced test acceptability were also considered to reflect an inadequate sample size rather than indicate effect modification. Similar trends were observed across odds ratios stratified by technician expertise for the same risk factor (refer to Table 8.8). On stratification by test reproducibility, a relationship between technician expertise and test acceptability was no longer detected among ATS reproducible tests (refer to Table 8.9). Similarly, after adjusting for acceptability of complementary spirograms, the previously established decreased "risk" of test failure associated with highly trained technicians pertained exclusively to non-acceptable complementary spirograms (refer to Table 8.10). Although the non-uniform stratum-specific odds ratios suggested that test reproducibility and complementary spirogram acceptability acted as effect modifiers of identified risk factors, estimates were considered unreliable because of small cell sizes.

8.2.5 Supplementary Analysis of Age and Gender

As mentioned in Chapter One, previous studies have identified both age and gender as "predictors" of test failure. To further verify the absence of their effect on test acceptability in the present study, mean differences in FEV_1 and FVC values between the best test and complementary spirogram's numeric results, in addition to the proportion of participants producing fewer than two acceptable curves or two non-reproducible curves, were examined according to age and gender. Results of the Primary and Secondary Data Sets are displayed in Tables 8.11 and 8.12, respectively. With the exception of one age category in each data set, no significant differences were detected.

As mentioned above, stratification created a series of categories in which the data were sparse and unbalanced. To determine whether such small sample sizes generated imprecise odds ratios and, in turn, erroneous inferences, potential determinants of test acceptability were re-examined using multivariate techniques.

8.3 Multivariate Analyses

In circumstances where stratification fails because of insufficient numbers, multivariate analysis is used to estimate the strength of associations while controlling for the effects of other factors (e.g., confounders) simultaneously (Hennekens, 1987). This technique involves construction of a mathematical model which relates a set of independent (explanatory) variables to a dependent variable (Kleinbaum, 1994). Selection of an appropriate multivariate model is based on the nature of the dependent variable. Since the present study examined a dichotomous outcome (best test spirogram: acceptable or non-acceptable), logistic regression analysis was performed. Logistic regression models the dependency of the probability of experiencing an outcome (e.g., performing an acceptable spirometric test) on a set of explanatory variables (covariates) through the relationship:

$$\ln[P/(1-P)] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_n X_n$$

P represents the probability of a dichotomous outcome and [P/(1-P)] represents the "odds" of experiencing a positive outcome (e.g., completion of an acceptable spirometric test). Both β_n and α are unknown parameters. Specifically, α refers to a constant term, the independent random error, and β_n is the coefficient for the n^th independent variable, X. The antilogarithm of each β coefficient represents the odds ratio associated with its corresponding variable (Hennekens, 1987).

Appropriate application of logistic regression methods requires that the data comply with a series of basic assumptions. Each case must represent an independent observation (i.e., the response of one case does not depend on that of another case). In addition, no multicollinearity should exist among independent variables. Multicollinearity occurs when independent variables in a model can be approximately determined by one or more of the other independent variables within the same model. Its presence can lead to unreliable estimated regression coefficients and inflated standard errors (Kleinbaum, 1994). The occurrence of complete separation (i.e., no overlap in the distribution of covariates between the two outcome groups) is also characterized by aberrant estimates and standard errors (Hosmer and Lemeshow, 1989).

Prior to performing logistic regression analysis in the present study, variables were screened graphically (through the construction of scatter plots) for multicollinearity. Assessment of the presence/absence of complete separation among variables involved inspection of standard errors. None of the variables in either data set demonstrated signs of these numerical problems.

8.3.1 Exact Logistic Regression

To estimate parameters of the logistic regression model, the conditional exact method of inference was employed. When unstratified data sets are small or non-normally distributed, asymptotic methods (based on maximizing the unconditional likelihood function) may fail to produce reliable results (Mehta and Patel, 1999). Maximum likelihood estimates are asymptotically normally distributed. Since bivariate analysis of both data sets generated sparse, unbalanced contingency tables comprised of cells with expected counts of less than 5[1], the use of standard, unconditional logistic regression was considered inappropriate. The unconditional maximum likelihood approach estimates all parameters (α and β_n) in the model. Subsequent hypothesis testing of these parameters is performed by computing Wald, likelihood ratio, or efficient scores statistics, all of which are based on a chi-square distribution. Conversely, the conditional exact method eliminates "nuisance parameters" by generating the permutation distributions of the sufficient statistics[2] for the parameters of interest, conditional on fixing the sufficient statistics of the remaining parameters at their observed values (Mehta and Patel, 1999).

The formula for the exact conditional likelihood estimate is an extension of the conditional likelihood function which reflects the probability of the observed data configuration relative to the probability of all possible configurations of the given data.

$$\frac{\prod_{l=1}^{m_1} \exp\left(\sum_{i=1}^{k} \beta_i X_{li}\right)}{\sum_{u}\left[\prod_{l=1}^{m_1} \exp\left(\sum_{i=1}^{k} \beta_i X_{uli}\right)\right]}$$

In this equation, m is the number of cases, X denotes a collection of variables for i ranging from 1 through k, β represents the coefficient corresponding to the covariate, X, and u is the number of possible combinations for selecting the l^{th} case. To calculate exact estimates, the exact inference is performed on each parameter, individually. By successively partitioning each parameter into two components (the first component is the

[1] When cells had expected counts of less than five, significance levels (i.e., p-values) based on Fisher's exact test were reported (SPSS, 9.0).

[2] The Sufficient statistic refers to the total number of positive outcomes for a covariate (factor) over all cases (Hosmer and Lemeshow, 1989).

parameter of interest and the second component is comprised of the remaining parameters in the model), the exact permutation distribution of the sufficient statistics for the parameter of interest, conditional on the sufficient statistics for the remaining parameters, is derived (Mehta and Patel, 1999).

8.3.2 Variable Specification

Multivariate analysis of both data sets considered all variables previously examined in the bivariate analysis (refer to Appendix 17).

With the exception of age and pack-years, all variables were dichotomous. These variables were assessed for collinearity. Observed correlation coefficients of 30.0% and 28.4% in the Primary Data Set and Secondary Data Set, respectively, indicated that age and pack-years were not highly correlated with each other. Logit plots were also constructed for each continuous variable. Neither displayed clear breaks or cut points. Consequently, age was dichotomized at 50 years based on its clinical significance. As discussed in Chapter One, prior studies have established a positive association between spirometric test failure and subjects over the age of 50 years. The variable, pack-years, was transformed into a categorical (dummy) variable comprised of three clinically relevant groups. Each group included an approximately equal number of participants. Both age and pack-years were "treated" similarly in both data sets.

A univariate analysis was completed on each variable. Those with an exact p-value of <0.25 were selected for the multivariate analysis (Hosmer and Lemeshow, 1989). Results of both data sets will be discussed in Section 8.4.

8.3.3 Model Building Strategies

To obtain the most parsimonious model (i.e., one that contained a minimal number of variables while accounting for the maximum amount of variance), variables were selected using a manual backward stepwise elimination approach (LogXact, 2.1). The initial model contained all of the variables. At each successive step in the model building process, the variable with the smallest significance level (i.e., largest p-value) was removed from the model. Variables previously excluded from the model were assessed for re-entry into the model. Both removal and re-entry were based on the exact conditional scores statistic.

Additionally, the likelihood ratio statistic, which measures twice the difference between the maximum log likelihood of two sequential models, along with its associated p-value for the chi-square distribution, was calculated for each step. This statistic tests the null hypothesis that coefficients of variables deleted are zero. Thus, differences in the log likelihood statistic were monitored across sequential steps for large, statistically significant changes upon variable removal (Hosmer and Lemeshow, 1989).

The "main effects" model derived for each data set was compared with that obtained using automated forward stepwise selection based on asymptotic inferences (specifically, the Likelihood Ratio Test) (SPSS, 9.0). Both techniques generated identical models. The computer-generated output detailing the model building steps is presented in Appendix 17.

Lastly, potential interaction terms, comprised of combinations of variables achieving a univariate significance of 0.25, were assessed by adding each term separately into the main effects model. Therefore, the final regression model retained significant covariates and interaction terms.

8.4 Results

i) The Primary Data Set

Findings from univariate analyses of variables in the Primary Data Set are presented in Table 8.13. Age, categorized pack-years, complementary spirogram acceptability, and prior spirometry experience each exhibited a univariate significance of $p < 0.25$ and, thus, appeared in the initial regression model. The resultant main effects model included a single variable, complementary spirogram acceptability (refer to Table 8.14). This model also became the final model since none of the potential interaction terms achieved statistical significance upon their entry (refer to Table 8.15).

ii) The Secondary Data Set

In the Secondary Data Set, the variables technician expertise, complementary spirogram acceptability, test reproducibility, and non-normal FEV_1 value produced a univariate significance of <0.25 (refer to Table 8.16). Of these variables, only technician expertise and complementary spirogram acceptability were retained in the main effects model (refer to Table 8.17). Addition of interaction terms generated non-significant findings (refer to Table 8.18). As observed in the Primary Data Set, the final model was identical to the main effects model.

To assess the fit of the final regression model (i.e., its effectiveness in describing the outcome, test acceptability), the Hosmer-Lemeshow Goodness-of-fit statistic was calculated. It is important to note that this statistic was only applied to models whose parameters were estimated by the unconditional maximum likelihood method. Goodness-of-fit techniques for models based on conditional exact methods are not currently available (Mehta and Patel, 1999).

No goodness-of-fit measure for the Primary Data Set's model could be calculated since the degrees of freedom were less than one. The Hosmer-Lemeshow Goodness-of-fit statistic for the Secondary Data Set was 2.43 (p-value of 0.30). The non-significant difference between the expected and observed probabilities indicated that the model was adequate (Hosmer and Lemeshow, 1989).

8.5 Conclusions

With respect to the Primary Data Set, bivariate and multivariate analyses of demographic, cardiopulmonary, and spirometry-related factors generated consistent findings. The single identified determinant of an acceptable best test spirogram was a similarly rated complementary spirogram. In the Secondary Data Set, bivariate methods related technician expertise, test reproducibility, and complementary spirogram acceptability to test acceptability whereas multivariate techniques identified technician expertise and complementary spirogram acceptability as the only contributing factors. Such differences may be a reflection of the data set's small sample size. Because the latter approach was based on exact (as opposed to asymptotic) inferences, its results were deemed more accurate.

Table 8.1 Acceptability of Best Test Spirograms According To Demographic, Cardiopulmonary Health, and Spirometry-Related Variables in the Primary Data Set

	Best Test Spirogram Quality			
Variable	Acceptable n (%)	Not Acceptable n (%)	Odds Ratio (95% CI)	p-value*
Demographic:				
Gender				
Female	17 (50.0)	33 (50.0)	1.000 (0.437-2.288)	> 0.999
Male	17 (50.0)	33 (50.0)		
Age				
Mean (sd)	39.19 (15.12)	43.84 (15.16)		0.114§
≥ 50	9 (26.5)	28 (42.4)	0.489 (0.198-1.208)	0.118
≤ 50	25 (73.5)	38 (57.6)		
Smoking Status				
Current	3 (8.8)	13 (19.7)	0.308 (0.079-1.202)	0.045
Past	6 (17.6)	20 (30.3)	0.400 (0.139-1.148)	0.084
Never	25 (73.6)	33 (50.0)	1.000 (Referent)	
Pack-years (sd)	4.40 (11.58)	9.16 (17.50)		0.108§
Cardiopulmonary:				
History of Shortness of Breath				
Yes	17 (50.0)	29 (43.9)	1.276 (0.557-2.925)	0.565
No	17 (50.0)	37 (56.1)		
History of Painful Breathing				
Yes	4 (11.8)	10 (15.2)	0.747 (0.216-2.584)	0.767†
No	30 (88.2)	56 (84.8)		
History of Wheeze or Asthma				
Yes	13 (38.2)	19 (28.8)	1.531 (0.640-3.667)	0.337
No	21 (61.8)	47 (71.2)		
History of Frequent Cough				
Yes	11 (32.4)	16 (24.2)	1.495 (0.600-3.723)	0.387
No	23 (67.6)	50 (75.8)		
Abnormal Chest X-Ray				
Yes	3 (8.8)	4 (6.1)	1.500 (0.316-7.123)	0.608†
No	31 (91.2)	62 (93.9)		
History of Lung Surgery				
Yes	0 (0.0)	1 (1.5)	-----‡	> 0.999†
No	34 (100.0)	65 (98.5)		
History of Heart Disease				
Yes	0 (0.0)	1 (1.5)	-----‡	> 0.999†
No	34 (100.0)	65 (98.5)		

Table 8.1 Continued

Variable	Best Test Spirogram Quality Acceptable n (%)	 Not Acceptable n (%)	Odds Ratio (95% CI)*	p-value*
Spirometry:				
Expertise of Technician Conducting Spirometry				
Highly Trained	18 (52.9)	35 (53.0)	0.996 (0.445-2.283)	0.993
Minimally Trained	16 (47.1)	31 (47.0)		
Percent of Predicted Lung Function Values From the Best Test Spirogram:				
FEV_1				
< lower limit of normal [¶]	3 (8.8)	3 (4.5)	2.032 (0.388-10.657)	0.393[†]
≥ lower limit of normal	31 (91.2)	63 (95.5)		
FVC			-----[‡]	0.305[†]
< lower limit of normal [¶]	0 (0.0)	2 (3.0)		
≥ lower limit of normal	34 (100)	64 (97.0)		
FEF_{25-75}			2.032 (0.388-10.657)	0.406[†]
< lower limit of normal[¶]	3 (8.8)	3 (4.5)		
≥ lower limit of normal	31 (91.2)	63 (95.5)		
Acceptability of Complementary Spirogram				
Acceptable	25 (73.5)	5 (7.6)	33.889 (10.329-111.192	<0.0001
Not Acceptable	9 (26.5)	61 (92.4)		
Reproducibility of the Best Test Spirogram's Results:				
Meets ATS Criteria	23 (67.6)	39 (59.1)	1.448 (0.606 - 3.455)	0.404
Does Not Meet ATS Criteria	11 (32.4)	27 (40.9)		
Completion of Spirometry Prior to Involvement in Present Study				
Yes	10 (29.4)	12 (18.2)	1.875 (0.628-5.469)	0.304
No	24 (70.6)	54 (81.8)		
Total	34	66		

*p-value based on chi-square statistic unless otherwise indicated

[†] p-value based on Fisher's exact test statistics

[§] p-value based on t-test statistic

[‡] No odds ratio calculated due to the presence of a zero cell

[¶] Lower limit of normal lung function values according to ATS criteria (ATS, 1994): FEV_1 - 80 % of predicted; FVC - 80 % of predicted; FEF_{25-75} - 50 % of predicted

Table 8.2 Acceptability of Best Test Spirograms According To Demographic and Spirometry-Related Variables in the Secondary Data Set

	Best Test Spirogram Quality			
	Acceptable	Not Acceptable		
Variable	n (%)	n (%)	Odds Ratio (95% CI)	p-value[*]
Demographic:				
Gender				
Female	60 (63.2)	62 (59.0)	1.189 (0.672-2.103)	0.552
Male	35 (36.8)	43 (41.0)		
Age				
Mean (sd)	35.95 (9.15)	36.41 (10.82)		0.746[§]
≥ 50	7 (7.4)	12 (11.4)	0.616 (0.232-1.637)	0.328[†]
< 50	88 (92.6)	93 (88.6)		
Smoking Status				
Current	13 (13.7)	18 (17.1)	0.756 (0.343-1.665)	0.486
Past	17 (17.9)	19 (18.1)	0.936 (0.448-1.957)	0.861
Never	65 (68.4)	68 (64.8)	1.000 (Referent)	
Pack-years (sd)	5.087 (11.53)	3.47 (8.10)		0.259[§]
Spirometry:				
Expertise of Technician Conducting Spirometry				
Highly Trained	61 (64.2)	39 (37.1)	3.036 (1.705-5.405)	< 0.001
Minimally Trained	34 (35.8)	66 (62.9)		
Percent of Predicted Lung Function Values From the Best Test Spirogram:				
FEV_1				
< lower limit of normal[¢]	2 (2.1)	7 (6.7)	0.301 (0.061-1.487)	0.175[†]
≥ lower limit of normal	93 (97.9)	98 (93.3)		
FVC				
< lower limit of normal[¢]	1 (1.1)	2 (1.9)	0.548 (0.049- 6.141)	> 0.999[†]
≥ lower limit of normal	94 (98.9)	103 (98.1)		
FEF_{25-75}				
< lower limit of normal[¢]	7 (7.4)	13 (12.4)	0.563 (0.215-1.476)	0.238[†]
≥ lower limit of normal	88 (92.6)	92 (87.6)		
Acceptability of Complementary Spirogram				
Acceptable	69 (72.6)	31 (29.5)	6.335 (3.422-11.727)	< 0.001
Not Acceptable	26 (27.4)	74 (70.5)		
Reproducibility of the Best Test Spirogram's Results				
Meets ATS Criteria	74 (77.9)	61 (58.1)	2.542 (1.367- 4.728)	0.003
Does Not Meet ATS Criteria	21 (22.1)	44 (41.9)		
Total	95	105		

[*] p-value based on chi-square statistic, unless otherwise indicated

[†] p-value based on Fisher's exact test statistics

[§] p-value based on t-test statistic

[¢] Lower limit of normal lung function values according to ATS criteria (ATS, 1994): FEV_1 - 80 % of predicted; FVC - 80 % of predicted; FEF_{25-75} - 50 % of predicted

Table 8.3 Acceptability of Best Test Spirograms According To Selected Variables After Stratification By Age Dichotomized At 50 Years in the Primary Data Set

	Best Test Spirogram Quality			
	Acceptable	Not Acceptable		
Variable	n	n	Odds Ratio (95% CI)	p-value[*]
Demographic Related:				
Gender				
Age ≥50				
Female	4	15	1.442 (0.319 - 6.529)	>0.887
Male	5	13		
Age <50				
Female	13	19	1.083 (0.395 - 2.974)	0.718
Male	12	19		
Crude Odds Ratio:			1.000 (0.437 - 2.288)	
Mantel-Haenszel Odds Ratio:			1.184 (0.512 - 7.738)	
Mantel-Haenszel χ^2 p-value:			0.858	
Spirometry Related:				
Reproducibility of the Best Test Spirogram's Results				
Age ≥50				
Meets ATS Criteria	7	18	1.944 (0.337 - 11.204)	0.687[†]
Does Not Meet ATS Criteria	2	10		
Age <50				
Meets ATS Criteria	16	21	1.439 (0.510 - 4.060)	0.491
Does Not Meet ATS Criteria	9	17		
Crude Odds Ratio:			1.448 (0.606 - 3.455)	
Mantel-Haenszel Odds Ratio:			1.563 (0.642 - 3.803)	
Mantel-Haenszel χ^2 p-value:			0.448	
Expertise Level of Technician Conducting Spirometric Test				
Age ≥50				
Highly Trained	6	18	1.111 (0.227 - 5.432)	>0.999[†]
Minimally Trained	3	10		
Age <50				
Highly Trained	12	17	1.140 (0.414 - 3.138)	0.779
Minimally Trained	13	21		
Crude Odds Ratio:			1.004 (0.438 - 2.248)	
Mantel-Haenszel Odds Ratio:			1.132 (0.482 - 2.657)	
Mantel-Haenszel χ^2 p-value:			0.947	

Table 8.3 Continued

Variable	Best Test Spirogram Quality Acceptable n	 Not Acceptable n	Odds Ratio (95% CI)	p-value*
Acceptability of Complementary Spirogram				
Age ≥50				
Acceptable	6	1	54.000 (4.754-613.325)	
Not Acceptable	3	27		<.0001⁺
Age <50				
Acceptable	19	4	26.916 (6.744 - 107.432)	
Not Acceptable	6	33		<0.0001
		Crude Odds Ratio:	33.889 (10.329-111.192)	
		Mantel-Haenszel Odds Ratio:	31.669 (9.523 - 105.314)	
		Mantel-Haenszel χ^2 p-value:	< 0.0001	

* p-value based on chi-square statistic, unless otherwise indicated

⁺ p-value based on Fisher's Exact Test Statistics

Table 8.4 Acceptability of Best Test Spirograms According To Selected Variables After Stratification By Expertise Level of the Technician Administering the Spirometric Test in the Primary Data Set

	Best Test Spirogram Quality			
	Acceptable	Not Acceptable		
Variable	n	n	Odds Ratio (95% CI)	p-value*
Demographic Related:				
Gender				
Highly Trained Technician				
Female	9	17	1.059 (0.340 - 3.301)	>0.999
Male	9	18		
Minimally Trained Technician				
Female	9	15	1.371 (0.408 - 4.614)	0.609
Male	7	16		
Crude Odds Ratio:			1.000 (0.437 - 2.288)	
Mantel-Haenszel Odds Ratio:			1.195 (0.522 - 2.737)	
Mantel-Haenszel χ^2 p-value:			0.835	
Age				
Highly Trained Technician				
≥ 50	6	18	0.472 (0.145 - 1.542)	0.210
< 50	12	17		
Minimally Trained Technician				
≥ 50	3	10	0.485 (0.112 - 2.095)	0.494[†]
< 50	13	21		
Crude Odds Ratio:			0.489 (0.198 - 1.208)	
Mantel-Haenszel Odds Ratio:			0.477 (0.190 - 1.198)	
Mantel-Haenszel χ^2 p-value:			0.176	
Spirometry Related:				
Reproducibility of the Best Test Spirogram's Results				
Highly Trained Technician				
Meets ATS Criteria	13	18	2.456 (0.721 - 8.368)	0.146
Does Not Meet ATS Criteria	5	17		
Minimally Trained Technician				
Meets ATS Criteria	10	21	0.794 (0.225 - 2.802)	0.719
Does Not Meet ATS Criteria	6	10		
Crude Odds Ratio:			1.448 (0.606 - 3.455)	
Mantel-Haenszel Odds Ratio:			1.438 (0.607 - 3.407)	
Mantel-Haenszel χ^2 p-value:			0.540	

Table 8.4 Continued

Variable	Best Test Spirogram Quality Acceptable n	 Not Acceptable n	Odds Ratio (95% CI)	p-value*
Acceptability of the Complementary Spirogram				
Highly Trained Technician				
Acceptable	15	3	53.333 (9.610-296.001)	<0.0001
Not Acceptable	3	32		
Minimally Trained Technician				
Acceptable	10	2	24.167 (4.182-139.669)	<0.0001†
Not Acceptable	6	29		
		Crude Odds Ratio:	33.889 (10.329 -111.192)	
		Mantel-Haenszel Odds Ratio:	35.817 (10.531 -121.819)	
		Mantel-Haenszel χ^2 p-value:	<0.0001	

* p-value based on chi-square statistic, unless otherwise indicated

† p-value based Fisher's Exact Test Statistics

Table 8.5 Acceptability of Best Test Spirograms According To Selected Variables After Stratification By ATS Reproducibility of the Best Test Spirogram's Results in the Primary Data Set

	Best Test Spirogam Quality			
	Acceptable	Not Acceptable		
Variable	n	n	Odds Ratio (95% CI)	p-value*
Demographic Related: Gender				
Best Test Spirogram Meets ATS Reproducibility Criteria				
Female	13	22	1.005 (0.355 - 2.840)	0.953
Male	10	17		
Best Test Spirogram Does Not Meet ATS Reproducibility Criteria				
Female	5	10	1.417 (0.342 - 5.866)	0.722†
Male	6	17		
		Crude Odds Ratio:	1.000 (0.437 - 2.288)	
		Mantel-Haenszel Odds Ratio:	1.131 (0.489-2.619)	
		Mantel-Haenszel χ^2 p-value:	0.941	
Age				
Best Test Spirogram Meets ATS Reproducibility Criteria				
≥ 50	7	18	0.510 (0.172 - 1.516)	0.223
< 50	16	21		
Best Test Spirogram Does Not Meet ATS Reproducibility Criteria				
≥ 50	2	10	0.378 (0.068 - 2.109)	0.444†
< 50	9	17		
		Crude Odds Ratio:	0.487 (0.198 - 1.208)	
		Mantel-Haenszel Odds Ratio:	0.468 (0.186 - 1.165)	
		Mantel-Haenszel χ^2 p-value:	0.157	
Spirometry Related: Expertise Level of Technician Conducting Spirometric Test				
Best Test Spirogram Meets ATS Reproducibility Criteria				
Highly Trained	13	18	1.517 (0.538 - 4.279)	0.430
Minimally Trained	10	21		
Best Test Spirogram Does Not Meet ATS Reproducibility Criteria				
Highly Trained	5	17	0.490 (0.118 - 2.030)	0.471†
Minimally Trained	6	10		
		Crude Odds Ratio:	1.004 (0.438 - 2.248)	
		Mantel-Haenszel Odds Ratio:	1.024 (0.449 - 2.334)	
		Mantel-Haenszel χ^2 p-value:	0.877	

Table 8.5 Continued

	Best Test Spirogram Quality			
	Acceptable	Not Acceptable		
Variable	n	n	Odds Ration (95% CI)	p-value*
Acceptability of Complementary Spirogram				
Best Test Spirogram Meets ATS Reproducibility Criteria				
Acceptable	18	4	31.500 (7.520-131.948)	<0.0001
Not Acceptable	5	35		
Best Test Spirogram Does Not Meet ATS Reproducibility Criteria				
Acceptable	7	1	45.500 (4.362- 474.646)	<0.0001†
Not Acceptable	4	26		
		Crude Odds Ratio:	33.889 (10.329-111.192)	
		Mantel-Haenszel Odds Ratio:	34.944 (10.312-118.414)	
		Mantel-Haenszel χ^2 p-value:	<0.0001	

* p-value based on chi-square statistic, unless otherwise indicated

† p-value based on Fisher's Exact Test

Table 8.6 Acceptability of Best Test Spirograms According To Selected Variables After Stratification By Acceptability of Complementary Spirograms in the Primary Data Set

	Best Test Spirogram Quality			
Variable	Acceptable n	Not Acceptable n	Odds Ratio (95% CI)	p-value*
Demographic Related:				
Gender				
Acceptable Complementary Spirogram				
Female	13	3	0.722 (0.102 - 5.095)	>0.999[†]
Male	12	2		
Not Acceptable Complementary Spirogram				
Female	5	29	1.379 (0.338 - 5.636)	0.731[†]
Male	4	32		
		Crude Odds Ratio:	1.000 (0.437 - 2.288)	
		Mantel-Haenszel Odds Ratio:	1.101 (0.356 - 3.419)	
		Mantel-Haenszel χ^2 p-value:	0.907	
Age				
Acceptable Complementary Spirogram				
≥ 50	6	1	1.263 (0.117 - 13.591)	0.999[†]
< 50	19	4		
Not Acceptable Complementary Spirogam				
≥ 50	3	27	0.630 (0.144 - 2.752)	0.723[†]
< 50	6	34		
		Crude Odds Ratio:	0.489 (0.198 - 1.202)	
		Mantel-Haenszel Odds Ratio:	0.766 (0.226 - 2.597)	
		Mantel-Haenszel χ^2 p-value:	0.624	
Spirometry Related:				
Expertise Level of Technician Conducting Spirometric Test				
Acceptable Complementary Spirogram				
Highly Trained	15	3	1.000 (0.141 - 7.099)	>0.999[†]
Minimally Trained	10	2		
Not Acceptable Complementary Spirogram				
Highly Trained	3	32	0.453 (0.104 - 1.979)	0.477[†]
Minimally Trained	6	29		
		Crude Odds Ratio:	1.004 (0.438 - 2.248)	
		Mantel-Haenszel Odds Ratio:	0.599 (0.186 - 1.909)	
		Mantel-Haenszel χ^2 p-value:	0.565	

Table 8.6 Continued

	Best Test Spirogram Quality			
	Acceptable	Not Acceptable		
Variable	n	n	Odds Ratio (95% CI)	p-value*
Reproducibility of the Best Test Spirogram's Results				
Acceptable Complementary Spirogram				
Meets ATS Criteria	18	4	0.643 (0.061 - 6.800)	>0.999[+]
Does Not Meet ATS Criteria	7	1		
Not Acceptable Complementary Spirogram				
Meets ATS Criteria	5	35	0.929 (0.227 - 3.801)	>0.999[+]
Does Not Meet ATS Criteria	4	26		
		Crude Odds Ratio:	1.448 (0.606 - 3.455)	
		Mantel-Haenszel Odds Ratio:	0.838 (0.253 - 2.776)	
		Mantel-Haenszel χ^2 p-value:	0.989	

* p-value based on chi-square statistic, unless otherwise indicated

[+] p-value based on Fisher's Exact Test

Table 8.7 Acceptability of Best Test Spirograms According To Selected Variables After Stratification By Age Dichotomized at 50 years in the Secondary Data Set

	Best Test Spirogram Quality			
	Acceptable	Not Acceptable		
Variable	n	n	Odds Ratio (95% CI)	p-value*
Demographic Related:				
Gender				
Age ≥50				
Female	4	6	1.333 (0.204-8.708)	> 0.999†
Male	3	6		
Age <50				
Female	56	56	1.156 (0.634-2.109)	0.636
Male	32	37		
		Crude Odds Ratio:	1.189 (0.672-2.103)	
		Mantel-Haenszel Odds Ratio:	1.172 (0.661-2.076)	
		Mantel-Haenszel χ^2 p-value:	0.693	
Spirometry Related:				
Reproducibility of the Best Test Spirogram's Results				
Age ≥50				
Meets ATS Criteria	6	7	4.286 (0.386-47.625)	0.333†
Does Not Meet ATS Criteria	1	5		
Age <50				
Meets ATS Criteria	68	54	2.456 (1.286-4.688)	0.006
Does Not Meet ATS Criteria	20	39		
		Crude Odds Ratio:	2.542 (1.367-4.728)	
		Mantel-Haenszel Odds Ratio:	2.562 (1.374-4.777)	
		Mantel-Haenszel χ^2 p-value:	0.005	
Expertise Level of Technician Conducting Spirometric Test				
Age ≥50				
Highly Trained	6	2	30.000 (2.217-405.98)	0.003†
Minimally Trained	1	10		
Age <50				
Highly Trained	55	37	2.523 (1.386-4.591)	0.002
Minimally Trained	33	56		
		Crude Odds Ratio:	3.036 (1.705-5.405)	
		Mantel-Haenszel Odds Ratio:	2.945 (1.663-5.214)	
		Mantel-Haenszel χ^2 p-value:	< 0.0001	

Table 8.7 Continued

Variable	Best Test Spirogram Quality Acceptable n	 Not Acceptable n	Odds Ratio (95% CI)	p-value*
Acceptability of Complementary Spirogram				
Age ≥50				
Acceptable	6	2	30.000 (2.217-405.98)	0.006[†]
Not Acceptable	1	10		
Age <50				
Acceptable	63	29	5.561 (2.938-10.527)	< 0.0001
Not Acceptable	25	64		
		Crude Odds Ratio:	6.335 (3.422-11.727)	
		Mantel-Haenszel Odds Ratio:	6.187 (3.350-11.428)	
		Mantel-Haenszel χ^2 p-value:	< 0.0001	

* p-value based on chi-square statistic, unless otherwise indicated

[†] p-value based on Fisher's Exact Test Statistics

Table 8.8 Acceptability of Best Test Spirograms According To Selected Variables After Stratification By Expertise Level of the Technician Administering the Spirometric Test in the Secondary Data Set

	Best Test Spirogram Quality			
	Acceptable	Not Acceptable		
Variable	n	n	Odds Ratio (95% CI)	p-value*
Demographic Related:				
Gender				
Highly Trained Technician				
Female	40	23	1.325 (0.579-3.034)	0.505
Male	21	16		
Minimally Trained Technician				
Female	20	39	0.989 (0.427-2.293)	0.979
Male	14	27		
		Crude Odds Ratio:	1.189 (0.672-2.103)	
		Mantel-Haenszel Odds Ratio:	1.147 (0.635-2.070)	
		Mantel-Haenszel χ^2 p-value:	0.762	
Age				
Highly Trained Technician				
≥ 50	6	2	2.018 (0.386-10.547)	0.477†
< 50	55	37		
Minimally Trained Technician				
≥ 50	1	10	0.170 (0.021-1.386)	0.092†
< 50	33	56		
		Crude Odds Ratio:	0.616 (0.232-1.637)	
		Mantel-Haenszel Odds Ratio:	0.632 (0.219-1.825)	
		Mantel-Haenszel χ^2 p-value:	0.575	
Spirometry Related:				
Reproducibility of the Best Test Spirogram's Results				
Highly Trained Technician				
Meets ATS Criteria	48	32	0.808 (0.291-2.244)	0.682
Does Not Meet ATS Criteria	13	7		
Minimally Trained Technician			4.147 (1.637-10.506)	0.002
Meets ATS Criteria	26	29		
Does Not Meet ATS Criteria	8	37		
		Crude Odds Ratio:	2.542 (1.367-4.728)	
		Mantel-Haenszel Odds Ratio:	2.003 (1.051-3.819)	
		Mantel-Haenszel χ^2 p-value:	0.051	

Table 8.8 Continued

Variable	Best Test Spirogram Quality Acceptable n	 Not Acceptable n	Odds Ratio (95% CI)	p-value*
Acceptability of the Complementary Spirogram				
Highly Trained Technician			3.947 (1.688-9.231)	0.001
Acceptable	42	14		
Not Acceptable	19	25		
Minimally Trained Technician			11.118 (4.099-30.152)	<0.0001
Acceptable	27	17		
Not Acceptable	7	49		
Crude Odds Ratio:			6.335 (3.422-11.727)	
Mantel-Haenszel Odds Ratio:			6.164 (3.261-11.648)	
Mantel-Haenszel χ^2 p-value:			<0.0001	

* p-value based on chi-square statistic, unless otherwise indicated

† p-value based on Fisher's Exact Test Statistics

Table 8.9 Acceptability of Best Test Spirograms According To Selected Variables After Stratification By ATS Reproducibility of the Best Test Spirogram's Results in the Secondary Data Set

	Best Test Spirogram Quality			
	Acceptable	Not Acceptable		
Variable	n	n	Odds Ratio (95% CI)	p-value*
Demographic Related: Gender				
Best Test Spirogram Meets ATS Reproducibility Criteria				
Female	48	39	1.041 (0.513-2.113)	0.911
Male	26	22		
Best Test Spirogram Does Not Meet ATS Reproducibility Criteria				
Female	12	23	1.217 (0.427-3.470)	0.713
Male	9	21		
	Crude Odds Ratio:		1.189 (0.672-2.103)	
	Mantel-Haenszel Odds Ratio:		1.094 (0.609-1.965)	
	Mantel-Haenszel χ^2 p-value:		0.881	
Age				
Best Test Spirogram Meets ATS Reproducibility Criteria				
≥ 50	6	7	0.681 (0.216-2.144)	0.509
< 50	68	54		
Best Test Spirogram Does Not Meet ATS Reproducibility Criteria				
≥ 50	1	5	0.390 (0.043-3.568)	0.655[†]
< 50	20	39		
	Crude Odds Ratio:		0.616 (0.232-1.637)	
	Mantel-Haenszel Odds Ratio:		0.592 (0.217-1.621)	
	Mantel-Haenszel χ^2 p-value:		0.442	
Spirometry Related: Expertise Level of Technician Conducting Spirometric Test				
Best Test Spirogram Meets ATS Reproducibility Criteria				
Highly Trained	26	29	1.673 (0.837-3.346)	0.144
Minimally Trained	48	32		
Best Test Spirogram Does Not Meet ATS Reproducibility Criteria				
Highly Trained	8	37	8.589 (2.600-28.378)	< 0.0001
Minimally Trained	13	7		
	Crude Odds Ratio:		3.036 (1.705-5.405)	
	Mantel-Haenszel Odds Ratio:		2.521 (1.406-4.521)	
	Mantel-Haenszel χ^2 p-value:		0.002	

Table 8.9 Continued

Variable	Best Test Spirogram Quality Acceptable n	 Not Acceptable n	Odds Ratio (95% CI)	p-value*
Acceptability of Complementary Spirogram				
Best Test Spirogram Meets ATS Reproducibility Criteria				
Acceptable	54	19	5.968 (2.830-12.588)	< 0.0001
Not Acceptable	20	42		
Best Test Spirogram Does Not Meet ATS Reproducibility Criteria				
Acceptable	15	12	6.667 (2.098-21.183	0.001
Not Acceptable	6	32		
Crude Odds Ratio:			6.335 (0.185-0.586)	
Mantel-Haenszel Odds Ratio:			6.166 (3.294-11.541)	
Mantel-Haenszel χ^2 p-value:			0.914	

* p-value based on chi-square statistic, unless otherwise indicated

† p-value based on Fisher's Exact Test

Table 8.10 Acceptability of Best Test Spirograms According To Selected Variables After Stratification By Acceptability of the Complementary Spirogram in the Secondary Data Set

	Best Test Spirogram Quality			
Variable	Acceptable n	Not Acceptable n	Odds Ratio (95% CI)	p-value*
Demographic Related:				
Gender				
Acceptable Complementary Spirogram				
Female	42	16	1.458 (0.621-3.427)	0.386
Male	27	15		
Not Acceptable Complementary Spirogram				
Female	18	46	1.370 (0.526-3.563)	0.518
Male	8	28		
		Crude Odds Ratio:	1.189 (0.672-2.103)	
		Mantel-Haenszel Odds Ratio:	1.418 (0.749-2.682)	
		Mantel-Haenszel χ^2 p-value:	0.363	
Age				
Acceptable Complementary Spirogram				
≥ 50	6	2	1.381 (0.26-7.260)	> 0.999†
< 50	63	29		
Not Acceptable Complementary Spirogam				
≥ 50	1	10	0.256 (0.031-2.105)	0.280†
< 50	25	64		
		Crude Odds Ratio:	0.616 (0.232-1.637)	
		Mantel-Haenszel Odds Ratio:	0.633 (0.201-1.990)	
		Mantel-Haenszel χ^2 p-value:	0.638	
Spirometry Related:				
Expertise Level of Technician Conducting Spirometric Test				
Acceptable Complementary Spirogram				
Highly Trained	42	14	1.889 (0.802-4.449)	0.143
Minimally Trained	27	17		
Not Acceptable Complementary Spirogram				
Highly Trained	19	25	5.320 (1.974-14.338)	0.001
Minimally Trained	7	49		
		Crude Odds Ratio:	3.036 (1.705-5.405)	
		Mantel-Haenszel Odds Ratio:	2.975 (1.575-5.618)	
		Mantel-Haenszel χ^2 p-value:	0.001	

Table 8.10 Continued

Variable	Best Test Spirogram Quality Acceptable n	 Not Acceptable n	Odds Ratio (95% CI)	p-value*
Reproducibility of the Best Test Spirogram's Results				
Acceptable Complementary Spirogram				
Meets ATS Criteria	20	42	2.274 (0.905-5.715)	0.077
Does Not Meet ATS Criteria	6	32		
Not Acceptable Complementary Spirogram				
Meets ATS Criteria	54	19	2.540 (0.914-7.055)	0.068
Does Not Meet ATS Criteria	15	12		
		Crude Odds Ratio:	2.542 (1.367-4.728)	
		Mantel-Haenszel Odds Ratio:	2.399 (1.210-4.755)	
		Mantel-Haenszel χ^2 p-value:	0.018	

* p-value based on chi-square statistic, unless otherwise indicated

' p-value based on Fisher's Exact Test

Table 8.11 FVC and FEV_1 Mean Differences Between Best Test and Complementary Spirograms, Number of Test Sessions With Fewer Than Two Acceptable Curves, and Number of Test Sessions Failing ATS Reproducibility Criteria According to Participant Age and Gender in the Primary Data Set

									< 2 Acceptable Curves			Non-Reproducible Curves		
		Number of	ΔFVC (ml)			ΔFEV_1 (ml)								
Age (years)	Gender	Participants	Mean	SEM	p-value†	Mean	SEM	p-value†	N	%	p-value*	N	%	p-value*
≤ 24	Male	6	188.33	56.06	0.081	173.33	57.60	0.216	4	66.67	> 0.999	3	50.00	0.617
	Female	13	66.15	10.65		90.00	17.39		2	15.38		4	30.76	
25 - 29	Male	6	158.33	40.37	0.034	115.00	45.22	0.272	3	50.00	0.545	3	50.00	0.812
	Female	5	42.00	13.93		52.00	22.00		1	20.00		0	0.00	
30 - 34	Male	4	122.50	65.37	-----	140.00	52.44	-----	4	100.00	-----	2	50.00	> 0.999
	Female	1	80.00	-----		110.00	-----		1	100.00		0	0.00	
35 - 39	Male	5	132.00	33.97	0.774	138.00	26.34	0.581	3	60.00	> 0.999	3	60.00	> 0.999
	Female	2	155.00	95.00		105.00	65.00		1	50.00		1	50.00	
40 - 44	Male	6	110.00	44.80	0.778	161.67	70.78	0.411	5	83.33	> 0.999	3	50.00	> 0.999
	Female	6	123.83	16.77		95.00	27.54		5	83.33		3	50.00	
45 - 49	Male	4	112.50	23.50	0.642	62.50	13.15	0.022	1	25.00	0.048	1	25.00	0.524
	Female	5	156.00	76.92		164.00	29.09		5	100.00		3	60.00	
50 - 54	Male	7	98.57	29.88	0.564	95.71	22.35	0.724	6	85.71	0.467	3	42.90	0.608
	Female	8	128.75	39.62		83.75	24.05		8	100.00		2	25.00	
55 - 59	Male	5	174.00	54.55	0.337	232.00	132.72	0.377	5	100.00	0.444	3	60.00	0.524
	Female	4	100.24	41.43		85.00	46.64		3	75.00		1	25.00	
≥ 60	Male	7	67.14	29.25	0.830	45.71	18.88	0.362	6	85.71	0.266	1	14.30	> 0.999
	Female	6	78.33	43.39		83.33	36.85		3	50.00		5	83.30	

† p-value based on t test

* p-value on Fisher's Exact Test

Table 8.12 FVC and FEV_1 Mean Differences Between Best Test and Complementary Spirograms, Number of Test Sessions With Fewer Than Two Acceptable Curves, and Number of Test Sessions Failing ATS Reproducibility Criteria According to Participant Age and Gender in the Secondary Data Set

			ΔFVC (ml)			ΔFEV_1 (ml)			< 2 Acceptable Curves			Non-Reproducible Curves		
Age (years)	Gender	Number of Participants	Mean	SEM	p-value†	Mean	SEM	p-value†	N	%	p-value*	N	%	p-value*
≤ 24	Male	13	148.50	23.77	0.294	81.54	23.06	0.069	9	69.23	> 0.999	7	53.85	0.041
	Female	16	105.60	30.43		32.50	10.14		10	62.50		2	12.50	
25 - 29	Male	14	141.40	41.06	0.104	101.40	23.20	0.190	11	79.57	> 0.999	2	14.29	0.676
	Female	21	76.19	16.64		60.48	19.58		16	76.19		5	23.80	
30 - 34	Male	11	70.91	27.02	0.083	82.73	20.81	0.241	7	63.63	0.696	2	18.18	0.231
	Female	16	146.90	29.48		115.60	17.70		8	50.00		7	43.75	
35 - 39	Male	9	182.20	48.24	0.855	171.11	86.96	0.498	6	66.67	> 0.999	5	55.55	0.418
	Female	21	195.00	40.97		115.60	38.54		15	71.43		7	33.33	
40 - 44	Male	11	197.30	70.36	0.237	105.50	23.10	0.052	7	63.64	> 0.999	4	36.36	0.438
	Female	22	101.18	30.79		53.33	7.05		14	63.64		5	22.73	
45 - 49	Male	11	207.00	42.51	0.548	64.00	24.18	0.540	5	45.45	0.452	7	63.64	0.252
	Female	16	160.00	54.58		81.25	15.94		10	62.50		6	37.50	
≥ 50	Male	9	191.10	44.08	0.923	115.60	26.46	0.765	6	66.67	> 0.999	3	33.33	> 0.999
	Female	10	172.00	44.57		111.10	36.83		7	70.00		3	30.00	

† p-value based on t test

* p-value on Fisher's Exact Test

Table 8.13 Univariate Analysis for the Exact Logistic Regression Analysis of the Primary Data Set

Variable	Odds Ratio*	Score*	p-value*
Gender	1.000	0.000	1.000
Age ≥ 50 years	0.492	2.450	0.118
Categorized Pack Years			
(1) 1 < pack-years < 10	0.281	3.982	0.064
(2) ≥ 10 pack-years	0.589	0.893	0.434
History of Shortness of Breath	1.273	0.332	0.565
History of Painful Breathing	0.749	0.212	0.767
History of Wheeze or Asthma	1.525	0.911	0.371
History of Frequent Cough	1.488	0.741	0.477
Abnormal Chest X-Ray	1.494	0.261	0.687
Expertise of Spirometry Technician	0.997	0.000	1.000
< Lower Limit of Normal (80% of predicted) FEV_1 Value	2.012	0.721	0.661
< Lower Limit of Normal (80% of predicted) FVC Value	1.255	1.048	0.547
Acceptability of Complementary Spirogram	31.959	46.482	0.000
ATS Reproducibility of Best Test Spirogram's Numeric Values	1.442	0.697	0.4037
Prior Completion of Spirometric Test	1.875	1.633	0.213

* Based on Exact Methods [i.e., Exact (Conditional Scores) Test]

Table 8.14 Exact Logistic Regression Main Effects Model Derived from the Primary Data Set

Model: BESTACPT = ACPT2

Number of Observations: 100
Number of Groups: 2
Degrees of Freedom:1
Exact (Conditional Scores) Statistic: 46.869
P- value: 0.000
Hosmer and Lemeshow Goodness-of-Fit Test:
Chi-Square (< df): Not Calculated
Significance: Not Calculated

	INFERENCE	<------------PARAMETER ESTIMATION------------>			P-VALUE
TERM	TYPE	ODDS RATIO	S.E	95.0% CONF. INTERVAL	2*1_SIDED
ACPT2	Exact	31.959	NA *	9.209 - 136.375	0.000
CONST	Asymptotic	0.1475	NA	0.073 - 0.297	0.000

Definition of Coded Terms:
Dependent Variable (Outcome):
BESTACPT: Accpetability of the Best Test Spirogram (Outcome or Response Term)
Independent Variable (Covariate):
ACPT2: Acceptability of the Complementary Spirogram
Constant:
CONST: Constant term

* Standard Errors for β and corresponding odds ratios (e^{β}) cannot be derived from the permutation distribution of the sufficient statistic for β, upon which exact inference methods are based.

Table 8.15 Analysis of Potential Interaction Terms in the Main Effects Model for the Exact Logistic Regression Analysis of the Primary Data Set

Interaction Term (coded variables)	Odds Ratio*	Score*	p-value*
Pack years (1 year-10 years) by:			
Age ≥ 50 years	0.784	0.781	0.377
ATS Reproducibility of Spirometric Results	1.375	1.023	0.312
Acceptability of Complementary Spirogram	0.2000	2.089	0.148
Completion of Previous Spirometric Test	0.507	1.225	0.268
Age ≥ 50 years by:			
ATS Reproducibility of Spirometric Results	0.583	0.603	0.437
Acceptability of Complementary Spirogram	0.381	1.167	0.279
Completion of Previous Spirometric Test	0.467	0.603	0.437
ATS Reproducibility of Spirometric Results by:			
Acceptability of Complementary Spirogram	3.810	2.305	0.129
Completion of Previous Spirometric Test	2.848	1.580	0.209
Acceptability of Complementary Spirogram by:			
Completion of Previous Spirometric Test	2.526	0.639	0.425

* Based on Exact Methods [i.e., Exact (Conditional Scores) Test]

Table 8.16 Univariate Analysis for the Exact Logistic Regression Analysis of the Secondary Data Set

Variable	Odds Ratio*	Score*	p-value*
Gender	1.188	0.352	0.565
Age $\geq$ 50 years	0.618	0.9515	0.348
Categorized Pack Years			
(1) 1 < pack-years < 10	0.652	0.889	0.346
(2) $\geq$ 10 pack-years	1.232	0.284	0.594
Expertise of Spirometry Technician	3.019	14.54	0.000
< Lower Limit of Normal (80% of predicted) FEV_1 Value	0.303	2.403	0.175
< Lower Limit of Normal (80% of predicted) FVC Value	0.550	0.244	0.621
Acceptability of Complementary Spirogram	6.268	36.887	0.000
ATS Reproducibility of Best Test Spirogram's Numeric Values	2.530	8.868	0.004

* Based on Exact Methods [i.e., Exact (Conditional Scores) Test]

Table 8.17 Exact Logistic Regression Main Effects Model Derived from the Secondary Data Set

Model: BESTACPT = ACPT2 + TECH

Number of Observations: 200
Number of Groups: 4
Degrees of Freedom: 2
Exact (Conditional Scores) Statistic: 46.542
P- value: 0.000
Hosmer and Lemeshow Goodness-of-Fit Test:
Chi-Square (2df): 2.427
Significance: 0.297

	INFERENCE	<------------PARAMETER	ESTIMATION	------------>	P-VALUE
TERM	TYPE	ODDS RATIO	S.E*	95.0% CONF. INTERVAL	2*1_SIDED
ACPT2	Exact	3.000	NA	1.514 - 5.975	0.001
TECH	Exact	6.302	NA	3.155 - 12.443	0.000
CONST	Asymptotic	0.070	NA	0.022 - 0.201	0.000

Definition of Coded Terms:
Dependent Variable (Outcome or Response):
BESTACPT: Accpetability of the Best Test Spirogram
Independent Variables (Covariates):
ACPT2: Acceptability of the Complementary Spirogram
TECH: Expertise of the Spirometry Technician
Constant:
CONST: Constant term

* Standard Errors for β and corresponding odds ratios (e^{β}) cannot be derived from the permutation distribution of the sufficient statistic for β, upon which exact inference methods are based.

Table 8.18 Analysis of Potential Interaction Terms in the Main Effects Model for the Logistic Regression Analysis of the Secondary Data Set

Interaction Term	Odds Ratio*	Score*	p-value*
Expertise of Spirometry Technician by:			
Acceptability of Complementary Spirogram	0.355	2.400	0.121
ATS Reproducibility of Spirometric Results	1.300	1.182	0.277
< Lower Limit of Normal (80% of predicted) FEV_1 value	0.423	3.172	0.075
Acceptability of Complementary Spirogram by:			
ATS Reproducibility of Spirometric Results	2.001	2.062	0.151
<Lower Limit of Normal (80% of predicted) FEV_1 value	0.078	7.464	0.053
ATS Reproducibility of Spirometric Results by:			
<Lower Limit of Normal (80% of predicted) FEV_1 value	0.182	3.298	0.069

* Based on Exact Methods [i.e., Exact (Conditional Scores) Test]

CHAPTER NINE

CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS

In this chapter, results are summarized and compared with hypothesized statements. Potential explanations for observed discrepancies, in addition to a discussion of the study's strengths and limitations, are provided. A set of suggestions regarding the successful use of spirometry in epidemiologic research is also presented.

9.1 Summary of Results

Inter-rater reliability analyses (Chapters 4 and 5) did not indicate a definite relationship between the magnitude of agreement among raters interpreting spirogram acceptability and their level of respiratory expertise. Further, no clear correlation between strengthened concordance and spirograms produced by technicians with greater spirometry training and experience was exhibited.

Proportions of best tests, complete test sessions, and ATS-reproducible sessions deemed acceptable across each rater category did not appear to vary according to rater expertise (Chapter 6). In the Primary Data Set, approximately two-fifths of spirograms received "acceptable" ratings, regardless of expertise category. Equal proportions were derived from spirometric tests conducted by minimally and highly trained technicians suggesting that technician expertise did not influence raters' interpretations of test acceptability. Despite calculation of lower values, similar patterns were observed for both complete test sessions and those complying strictly with ATS Reproducibility Criteria. The Secondary Data Set produced different findings. While roughly half of the tracings were judged acceptable by Non-certified, Minimally Trained, Respiratory Research Assistants, Respiratory Epidemiologists and Pulmonary Specialists; a considerably lower proportion (one-fifth) was accepted by Certified, Respiratory Technicians. In addition, results from raters with professional respiratory expertise demonstrated a positive effect of increased technician expertise on the proportion of acceptable best test spirograms, and complete spirometry sessions.

No stable relationship between raters' respiratory expertise and the types of explanations provided for spirometric test failure was established (Chapter 7). While complete categories of raters defined curve patterns comprising the Primary Data Set differently, raters across all categories interpreted artifacts displayed on curves from the Secondary Data Set similarly.

Findings from multivariate analyses of potential determinants of spirometric test acceptability (demographic, cardiopulmonary, and spirometry-related) were data set-specific. Complementary spirogram acceptability was the only contributing factor identified in the Primary Data Set. With respect to the Secondary Data Set, technician expertise, in addition to complementary spirogram acceptability, was related to test acceptability. Although not confirmed, ATS reproducibility of corresponding spirometric test sessions appeared to modify the relationship between technician expertise and best test acceptability. Higher technician expertise correlated with increased test acceptability for spirograms from test sessions failing to comply with ATS reproducibility criteria.

9.2 Consistency with Initial Hypotheses

None of the study's findings confirmed hypotheses specified *a priori*. It was expected that the degree of agreement among raters would depend on the similarity of their respiratory expertise. Contrary to this hypothesis, results indicated that raters of a common expertise level did not agree most strongly with each other. Additionally, the anticipated positive effect of technician expertise on rater concordance was not confirmed. Neither data set demonstrated a consistent relationship between strengthened inter-rater reliability and spirograms produced by highly trained technicians. Further, the proportion of spirograms deemed acceptable by each rater category increased significantly with technician expertise in the Secondary Data Set, only. Lastly, it was hypothesized that certain characteristics of participants performing spirometric maneuvers would influence spirometric test acceptability. None of the recorded demographic or cardiopulmonary health-related factors proved to be significant. Additionally, participants who had completed spirometric tests prior to their involvement in the current study did not produce a greater proportion of acceptable tests, demonstrating the absence of the long term practice or learning effect. However, sample sizes were not sufficient to achieve adequate statistical power.

9.3 Agreement of Findings with Previous Studies of Spirometric Test Acceptability

Because review of the literature located only a few investigations that examined spirometric test quality, the extent to which findings from this study could be compared with those from previous research was limited.

The New Zealand study of spirometric test quality within a clinical setting (refer to Section1.5.2) reported no significant difference in the proportion of "correct" interpretations between "trained" and "untrained" primary care practitioners (Two pulmonologists, serving as gold standards, determined accuracy) (Eaton et al, 1999). This finding, which suggested that raters representing two distinct levels of respiratory expertise interpreted spirogram quality similarly, correlated with the present study's results. However, methodological differences between the two study designs precluded complete comparability of results. For example, participants performing spirometric maneuvers were patients who presented with clinical indications for spirometry [e.g., management of asthma, investigation of respiratory symptoms or COPD (chronic obstructive pulmonary disease)] rather than individuals from the general population. Consequently, the types of curve patterns presented to raters in each study were likely different. Additionally, the New Zealand Study did not formally address inter-rater reliability between practitioners or the two "gold standards".

Agreement among pulmonologists regarding spirometric interpretations was measured by

Quadrelli and co-workers (refer to Section 1.5.2.ii) (Quadrelli et al, 1996). Concordance reached a maximum value of 76%. The percentage of agreement between pulmonary specialists involved in the present study was calculated for each data set in order to facilitate a comparison (Recall that original statistical strategies for quantifying inter-rater reliability employed kappa which measures chance-corrected agreement as opposed to overall agreement). Similar values were obtained [76% ($\kappa = 0.50$) and 73% ($\kappa = 0.40$) in the Primary and Secondary Data Sets, respectively], supporting the validity of present study's findings.

Since no study examining the effect of technician training and experience on the degree of discordance among raters evaluating spirometric tests was located, results could not be assessed for consistency with previous findings. Further, none of the studies discussed in Chapter One quantified differences in the proportion of acceptable spirograms from test sessions conducted by certified and uncertified technicians. Instead, a positive "training effect" was established subsequent to observing (either retrospectively or prospectively) an increase in test acceptability as a single cohort of technicians acquired additional training and experience through a given time period (Enright et al, 1991; Hankinson and Moon Bang,1991; Kunzlie et al, 1995). Therefore, these results were not compared with those of the present cross-sectional study.

Contrary to the findings of both the NHANES III and Lung Health Study (refer to Section 1.6.2.i), increased test failure was not associated with either female participants or those over 49 years of age (Hankinson and Moon Bang, 1991; Enright et al, 1991). Prior studies have also related test failure to the presence of respiratory symptoms (Becklake, 1990). None of the cardiopulmonary health factors measured in the present study were identified as determinants of spirogram non-acceptability. However, results, although not statistically significant, suggested a "learning or practice effect" comparable to that observed for participants in The Lung Health Study. An *ad hoc* comparative analysis of variations in the proportion of acceptable spirograms between test sessions was performed to determine whether greater test acceptability was associated with participants' second spirometry sessions in the Primary Data Set. The order in which participants were tested by each technician was also considered (refer to Appendix 18). Regardless of technician order, the number of participants who achieved an acceptable first test but non-acceptable second test was lower than that of participants who produced a non-acceptable first test but acceptable second test (p-values based on McNemar's test >0.100).

9.4 Possible Explanations for Differences in Results of the Primary and Secondary Data Sets

The two data sets included equivalently trained spirometry technicians who tested comparable participant populations within a field setting. In addition, the number of consecutive test sessions administered during a given time period was similar. Raters interpreting spirometric results were common to both sets. However, each data set employed a different type (i.e., make and model) of portable pneumotach[1] spirometer that featured distinctive damping characteristics. To review, damping refers to an instrument's ability to accurately respond to dynamic changes in air flow (Clausen,1982). The degree of damping incorporated into each spirometer's pressure transducer is evidenced by the appearance of tracings generated. Spirograms comprising the Primary Data Set displayed smooth or "damped patterns" indicating that no "overshooting"of the output signal occurred. In contrast, curves collected in the Secondary Data Set exhibited "underdamped" jagged patterns which may be explained by an "overshooting" of the input signal. Raters' frequently ascribed such flow fluctuations to participant "coughing", "variable effort" or "turbulence"(refer back to Figures 7.1a-c and 7.2a-e). It is possible that inconsistencies between the Primary and Secondary Data Sets' findings reflected differences in the equipments' damping mechanisms.

[1] A pneumotach spirometer measures instantaneous air flow using a differential pressure device (Wanger, 1992).

9.5 Limitations of the Study

Several limitations of the study design were recognized.

9.5.1 Participant Selection Bias

All participants who received spirometry testing during the study's duration were volunteers. Archived spirograms in the Secondary Data Set also originated from participants who voluntarily entered a respiratory health research study. Therefore, both study populations (i.e., the Primary and Secondary) consisted exclusively of volunteers, equalizing any "volunteer effect". However, since participation was based on a self-selection process, the majority of individuals who received spirometry testing appeared healthy. Consequently, interpreters' assessments of the acceptability of abnormal lung function results could not be evaluated. Nevertheless, it is important to note that most epidemiologic investigations examine general populations. In clinical settings, interpreters encounter atypical curve patterns at a rate rarely observed in the field. Analysis of those curves may require a level of expertise different from that necessary for epidemiologic research. Thus, it is likely that results of the present study are not generalizable to a clinical environment

9.5.2 The Hawthorne Effect

It is possible that both raters and technicians "behaved" differently simply because they were advised of the study's objectives (Hawthorne effect) (Last, 1995). Although technicians who generated archived spirograms were aware that output would be periodically assessed for quality, their performance may have changed if they had been asked to conduct spirometry for the sole purpose of addressing the present study's objectives. However, all raters were blinded to the expertise level of technicians who generated each spirogram, minimizing the presence of the Hawthorne effect on observed relationships between technician expertise and raters' interpretations of test acceptability.

9.5.3 Sample Sizes of Raters, Technicians, and Participants

The low prevalence of certain participant characteristics (e.g., age over 50 and presence of cardiopulmonary health conditions) resulted in small cell numbers and, in turn, low statistical power. Therefore, it was not possible to achieve reliable odds ratios for certain potential predictors of spirometric test failure. Further research involving larger sample sizes would help to verify the present study's findings.

Only a small number of raters and technicians comprised expertise levels. Since results between data sets were discrepant, it became difficult to determine whether technicians or raters accurately represented the expertise of those with similar training and experience who were not involved in the study. A more complete assessment would require increasing the size of both groups.

9.5.4 Intra-rater Reliability

This study did not address intra-rater reliability for spirometric test interpretations. Such analyses measure discrepancies between repeated observations by the same rater. Since neither data set included more than one copy of a single curve, the reproducibility (i.e., reliability) of each rater's assessments could not be examined.

9.6 Implications and Recommendations Based on the Study's Findings

Results of this study indicate that the interpretation of spirometric test quality is neither simple nor straightforward. Inconsistencies noted between data sets and among expertise levels suggest the potential for further investigation of procedural and interpretation strategies.

As mentioned in Section 9.5.1, the type of pulmonary function training and experience necessary for the successful use and interpretation of spirometry in primary care settings may differ significantly from that needed to collect data in an epidemiologic study. Availability of supplementary laboratory equipment, the opportunity for comparison of "serial"curve patterns included within each patient's record, and access to direct professional supervisory advice are often conditions unique to the clinic environment. Importantly, all Certified, Respiratory Technicians who participated in the present study acquired their experience in pulmonary function laboratories. Respiratory epidemiologic field studies are often conducted without the direct assistance of a pulmonary specialist or access to a well-equipped pulmonary function laboratory. Further analysis of the quality of spirometry administered by both minimally and highly trained technicians in primary care practices, pulmonary function laboratories, and field settings could assist in determining the qualifications most applicable to each environment.

All previous studies indicating a positive effect of technician expertise on test acceptability provided continual supervisory feedback to technicians regarding spirogram quality (Enright et al, 1991). Although the present study did not offer the rater and technician an opportunity to communicate during or after test administration, such contact may have addressed many tracing ambiguities or discrepancies that were encountered. Nevertheless, if direct communication is not a viable option (as in the case of archived data), detailed notes written by the technician during test administration could accompany spirometric tracings. Such comments may facilitate a more accurate, retrospective interpretation of test quality. In the data collected for the present study, one rater indicated the preference for detailed technicians' notes when attempting to differentiate between a "cough" and "turbulence".

This study further confirmed the need for technicians to possess a complete knowledge of spirometric equipment complexities in order to distinguish and accurately note malfunctions at the time of testing. In epidemiology, this awareness becomes critical since field studies are often conducted in remote locations where access to supervisory expertise is limited.

Finally, results reiterate the importance of careful, current, and thorough review of tracings collected in field studies. Although the level of respiratory expertise required to accurately "rate" spirogram quality remains uncertain, the need for independent assessments from a series of raters for identification of questionable spirometric results was clearly established. A more comprehensive reference collection of sample tracings in the ATS standardization document might facilitate efficient, effective interpretation of curve artifacts and, thus, ensure optimal quality of lung function data in respiratory epidemiologic studies.

REFERENCES

Altman DG. *Practical Statistics for Medical Research.* London, UK: Chapman and Hall; 1991.

American Thoracic Society. Lung function testing: selection of reference values and interpretative strategies. *Am Rev Respir Dis* 1991; 144: 1202-1218.

American Thoracic Society. Snowbird workshop on standardization of spirometry. *Am Rev Respir Dis* 1987; 136: 1285-1298.

American Thoracic Society. Standardization of spirometry. *Am Rev Respir Dis* 1979; 119: 831-838.

American Thoracic Society. Standardization of spirometry - 1994 update. *Am J Respir Crit Care Med* 1995; 152: 1107-1136.

Becklake MR. Epidemiology of spirometric failure. *Br J Ind Med* 1990; 47: 73-74.

Becklake MR, White N. Sources of variation in spirometric measurements. Identifying the signal and dealing with noise. *Occupational Medicine* 1993; 8(2): 241-264.

Behringer E, Rees J, Davies J. Poorly performed lung function tests - The answer is not blowing in the wind. *South Afr Med J* 1991; 80: 313-314.

Bosse CG, Criner GJ. Using spirometry in the primary care office. A guide to technique and interpretation of results. *Spirometry* 1993; 93: 122-148.

Burrows B, Lebowitz MD, Camilli AE, Knudson RJ. Longitudinal changes in forced expiratory volume in one second in Adults. Methodologic considerations and findings in healthy nonsmokers. *Am Rev Respir Dis* 1986; 133: 974-980.

Camilli AE, Burrows B, Knudson RJ, Lyle SK, Lebowitz MC. Longitudinal changes in forced expiratory volume in one second in adults: effects of smoking and smoking cessation. *Am Rev Respir Dis* 1987; 135: 794-799.

Christiani D, Eisen EA, Qegman DH. Respiratory disease in Chinese cotton textile workers, I: Respiratory symptoms. *Scand J Work Environ Health* 1985; 125:56-59.

Chusid, E. Leslie, M.D. *The Selective and Comprehensive Testing of Adult Pulmonary Function.* Mount Kisco, New York: Futura Publishing; 1983.

Clausen, Jack L. *Pulmonary Function Testing Guidelines and Controversies (Equipment, Methods, and Normal Values)* New York, New York: Academic Press; 1982.

Crapo, Robert O., M.D. Pulmonary function testing. *The New England Journal of Medicine* 1994; 331: 25.

Crapo RO, Reed MG, Berlin SL, Morris AH. Automation of pulmonary function equipment. User beware! *Chest* 1986; 90:1-2.

Daniel WW. Biostatistics. *A Foundation for Analysis in the Health Sciences* (6th Ed). Toronto, ON: John Wiley & Sons; 1995.

Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* 1992; 11: 1511-1519.

Eaton T, Withy S, Garrett JE, Mercer J, Whitlock RML, Rea HH. Spirometry in primary care practice. The importance of quality assurance and the impact of spirometry workshops. *Chest* 1999; 116: 416-423.

Eisen EA. Standardizing spirometry: Problems and prospects. *Occup Med State Art Rev* 1987; 2: 213-226.

Eisen EA, Oliver LC, Christiani DC, Robins JM, Wegman DH. Effects of spirometry standards in two occupational cohorts. *Am Rev Respir Dis* 1985; 132: 120.

Enright PL, Johnson LJ, Connett JE. Spirometry in the Lung Health Study: methods and quality control. *Am Rev Respir Dis* 1991; 143: 1215-1223.

Ferris BG. Epidemiology standardization project: recommended standardized procedures for pulmonary function testing. *Am Rev Respir Dis* 1978; 118: 55-88.

Fleiss JL. *Statistical Methods for Rates and Proportions (2nd Ed).* Toronto, ON: John Wiley & Sons; 1981.

Gardner RM, Clausen JL, Epler G, Hankinson JL, Permutt S, Plummer AL. Pulmonary Function Laboratory Personnel Qualifications. *ATS News* 1983; 9:(3 - Summer)12-16.

Glindmeyer HW, Jones RN, Barkman HW, Weill H. Spirometry: quantitative test criteria and test acceptability. *Am Rev Respir Dis* 1987; 136: 449-452

Hankinson JL. Pulmonary function testing in the screening of workers: guidelines for instrumentation, performance, and interpretation. *Journal of Occupational Medicine* 1986; 28: 1081-92

Hankinson JL and Moon Bang K. Acceptability and reproducibility criteria of the American Thoracic Society as observed in a sample of the general population. *Am Rev Respir Dis* 1991; 143: 516-521.

Hennekens CH and Buring JE. *Epidemiology in Medicine.* Toronto, ON: Little, Brown and Company; 1987.

Hnatiuk O, Moores L, Loughney T, Torrington K. Evaluation of internists' spirometric interpretations. *J Gen Intern Med* 1996; 11: 204-208

Hosmer DW, Lemeshow S. *Applied Logistic Regression.* Toronto, ON: John Wiley and Sons; 1989.

Hughes DTD and Empey DW. *Lung Function for the Clinician.* London, Gr. Britain: Academic Press; 1981.

IUATLD. *International Union Against Tuberculosis and Lung Disease Bronchial Symptom Questionnaire* 1986; Paris: IUATLD.

Kellie SE, Attfield MD, Hankinson JL, Castellan RM. The ATS spirometry variability criteria: Associations with morbidity and mortality in an occupational exposed cohort of coal miners. *Am J Epidemiol* 1987; 125; 437-444.

Kleinbaum DG. *Logistic Regression. A Self-Learning Text.* New York, New York: Springer-Verlan; 1994.

Kraemer HC and Bloch DA. Kappa Coefficients in epidemiology: an appraisal of a reappraisal. *J Clin Epidemiol* 1989; 41: 959-968.

Krzyzanowski M, Jedrychowski W, Wysocki W. Significance of spirometry repeatability criteria for the assessment of lung function level and changes over 13 years. *Am Rev Respir Dis* 1988; 137: 256.

Kunzlie N, Ackermann-Liebrich U, Keller AP, Perruchoud AP, Schindler C, SAPALDIA team. Variability of FVC and FEV1 due to technician, team, device, and subject in an eight centre study: three quality control studies in SAPASDIA. *Eur Respir J* 1995; 8: 371-376.

Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-174.

Last JM. *A Dictionary of Epidemiology.* New York, New York: Oxford University Press; 1995.

Laszlo F, Sudlow MR. The work of the respiratory laboratory. *Measurement in Clinical Respiratory Physiology.* New York, New York: Academic Press; 1983.

Manolio TA, Weinmann GG, Buist AS, Furberg CD, Pinsky JL, Hurd SH. Pulmonary function testing in population-based studies. *Am J Respir Crit Care Med* 1997; 156: 1004-1010

Mantel N and Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI* 1959; 22:719 - 748.

McKay RT, Lockey JE. Pulmonary function testing: Guidelines for medical surveillance and epidemiologic studies. *Occup Med State Art Rev* 1991; 6: 43-57.

Mehta C, Patel N. *LogXact for Windows. User Manual.* Cambridge, MA: CYTEL Software Corporation; 1999.

Miller, A (Ed). *Pulmonary Function Tests, A Guide for the Student House Officer.* Orlando, Florida: Grune and Stratton; 1987.

Nelson SB, Gardner RM, Crapo RO, Jensen RL. Performance evaluation of contemporary spirometers. *Chest* 1990; 97: 288-297.

Norman GR. Measures of association for categorical data. *Biostatistics: the Bare Essentials.* St. Louis, MI: Mosby Year Book; 1994.

Norman GR, Streiner DL. *PDQ Statistics.* St. Louis, MI: Mosby Year Book; 1997.

Norusis MJ. *SPSS Regression Models 9.0.* Chicago, Il: SPSS Inc.; 1999.

Oliver LC, Eisen EA, Green RE, Sprince NL. Asbestos-related disease in railroad workers: a cross-sectional study. *Am Rev Respir Dis* 1985; 131:499-504.

Petty TL. The predictive value of spirometry: Identifying patients at risk for lung cancer in the primary care setting. *Post Graduate Medicine* 1997; 101(3): 128-140.

Posner KL, Sampson PD, Caplan RA, Ward RJ, Cheney FW. Measuring interrater reliability among multiple raters: an example of methods for nominal data. *Statistics in Medicine* 1990; 9: 1103-1115.

Quadrelli SA, Roncoroni AJ, Porcel G. Analysis of variability in interpretation of spirometric tests. *Respiration* 1996; 63: 131-136

Quanjer H, Tammeling GJ, Cotes JE, Pedersen OF, Peslin R, Yernault JC. Lung volumes and forced ventilatory flows. *Eur Respir* J 1993; 6, Suppl 16: 5-40.

Richardson DB. Respiratory effects of chronic hydrogen sulfide exposure. *Am J of Ind Med* 1995; 28: 99-108.

Rothman KJ, Greenland S. *Modern Epidemiology.* Philadelphia, PA: Lippincott-Raven Publishers; 1998.

Ruppel, Gregg L. Spirometry. *Respiratory Care Clinics of North America* 1997; 3: 155-181.

Russell NJ, Crichton NJ, Emerson PA, Morgan AD. Quantitative assessment of the value of spirometry. *Thorax* 1986; 41: 360-363.

Sly, M.R.. Mortality from asthma. *Journal of Allergy and Clinical Immunology* 1989; 84: 421-34.

SPSS *Graduate Pack 9.0 for Windows* (Microsoft Software) 1999.

Thompson-Dobo. (Personal correspondence with market firm, 1999).

Tockman MS, Comstock GW. Respiratory risk factors and mortality; longitudinal studies in Washington county, Maryland. *Am Rev Respir Dis* 1989; 140 (Suppl:56-63).

Walter S. E-mail correspondence, 1999.

Wanger, Jack. *Pulmonary Function Testing: A Practical Approach.* Baltimore, Maryland: Williams & Wilkins; 1992.

Wenzel SE, Larsen GL. Assessment of lung function: pulmonary function testing. *Allergy, Asthma and Immunology from Infancy to Adulthood.* Philadelphia, PA: W.B. Saunders; 1996.

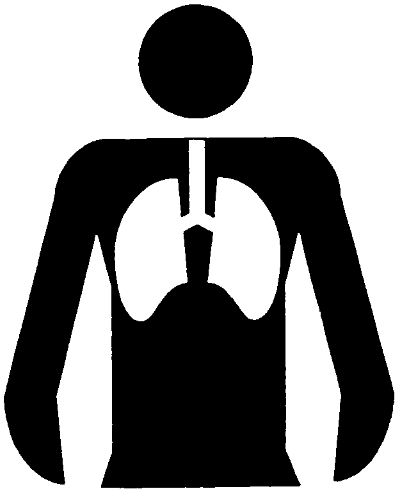
Westdorp EJ, Gratton MC, Watson WA. Emergency department interpretation of electrocardiograms. *Annals of Emergency Medicine* 1992; 21: 83-86.

Wise RA, Connett J, Kurnow K, Grill J, Johnson L, Kanner R , Enright P, Lung Health Study Group. Selection of spirometric measurements in a clinical trial, the Lung Health Study. *Am J Respir Crit Care Med* 1995; 151: 675-681.

Appendix 1:

Lung Health Awareness Clinic Advertisement

(Participant Recruitment for Establishment of Primary Data Set)



How Healthy Are Your Lungs?

The Alberta Asthma Centre is holding a **Lung Health Awareness Clinic**

Where? **Southgate Mall** in front of the Public Library entrance.

When? **Tuesday thru Saturday** (February 2 - 6, 1999)

Hours? Tuesday thru Friday: **10:00 am to 9:00 pm**

Saturday: **10:00 am to 5:00 pm**

In addition to information concerning lung health awareness,
free lung function testing will be conducted by trained technicians.

Testing Times: Tuesday thru Friday - 4:00 to 9:00 pm
Saturday - 10:00 am to 5:00 pm

Better Breathing - Better Living

ALBERTA ASTHMA CENTRE

Appendix 2:

Lung Health Clinic Field Setting

Display and Testing Centre

Southgate Mall, Edmonton, Alberta

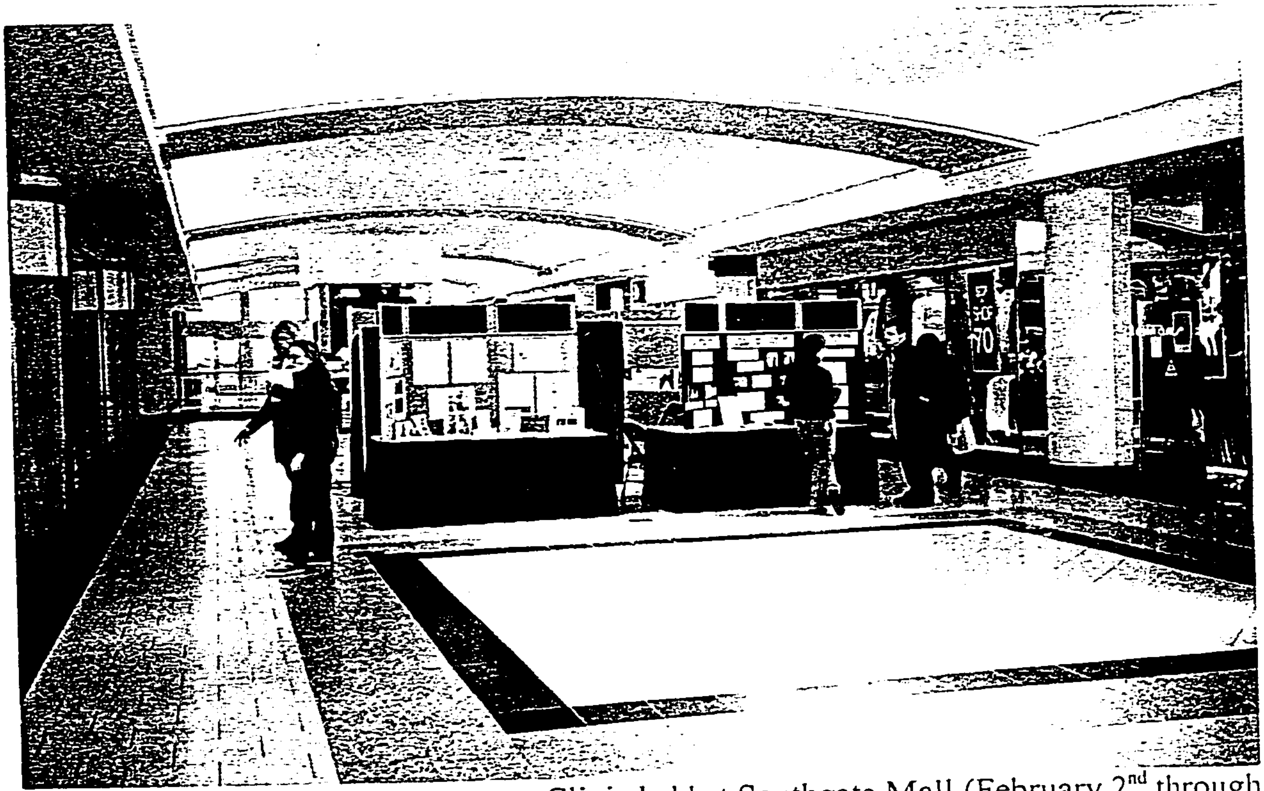


Figure 2a1. Lung Health Awareness Clinic held at Southgate Mall (February 2nd through February 6th, 1999) in Edmonton, Alberta.

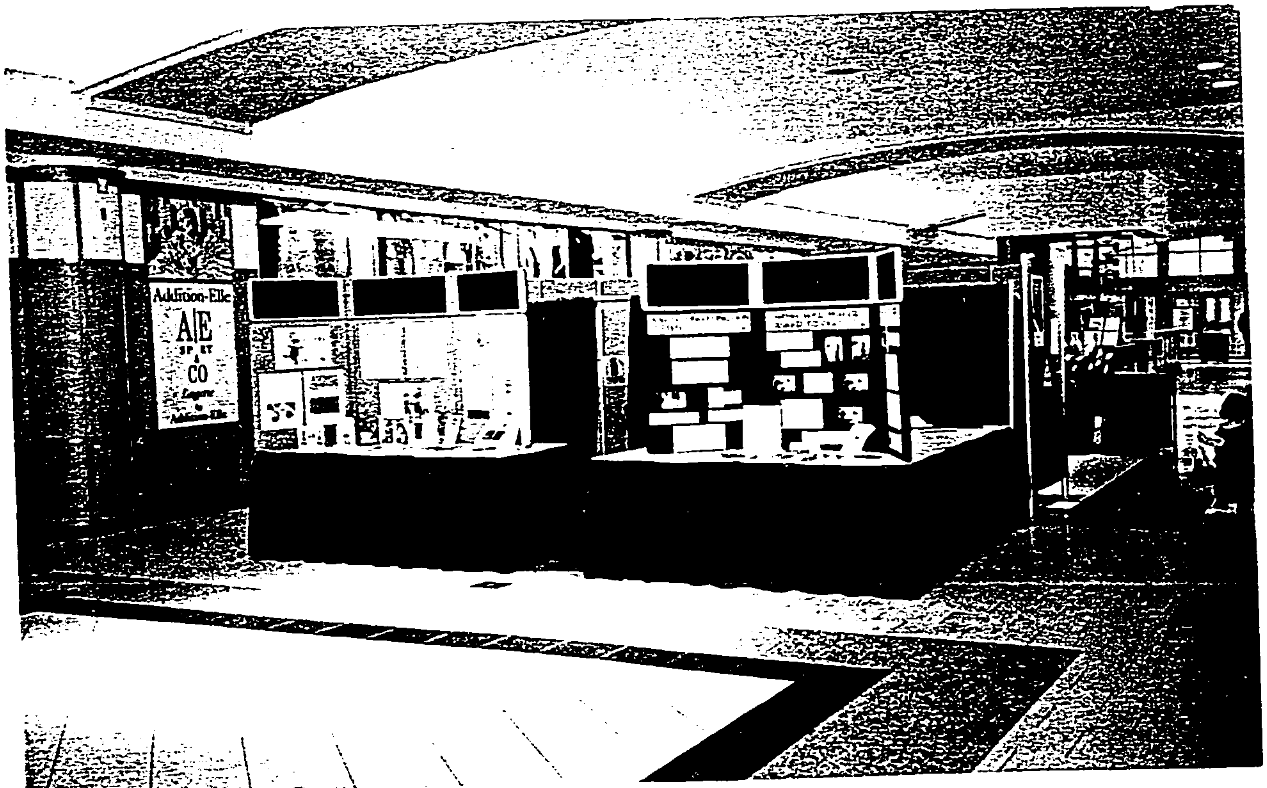


Figure 2a2. Asthma Centre and Study Display

Figure 2a3. Asthma Centre Display providing information on asthma and general pulmonary health

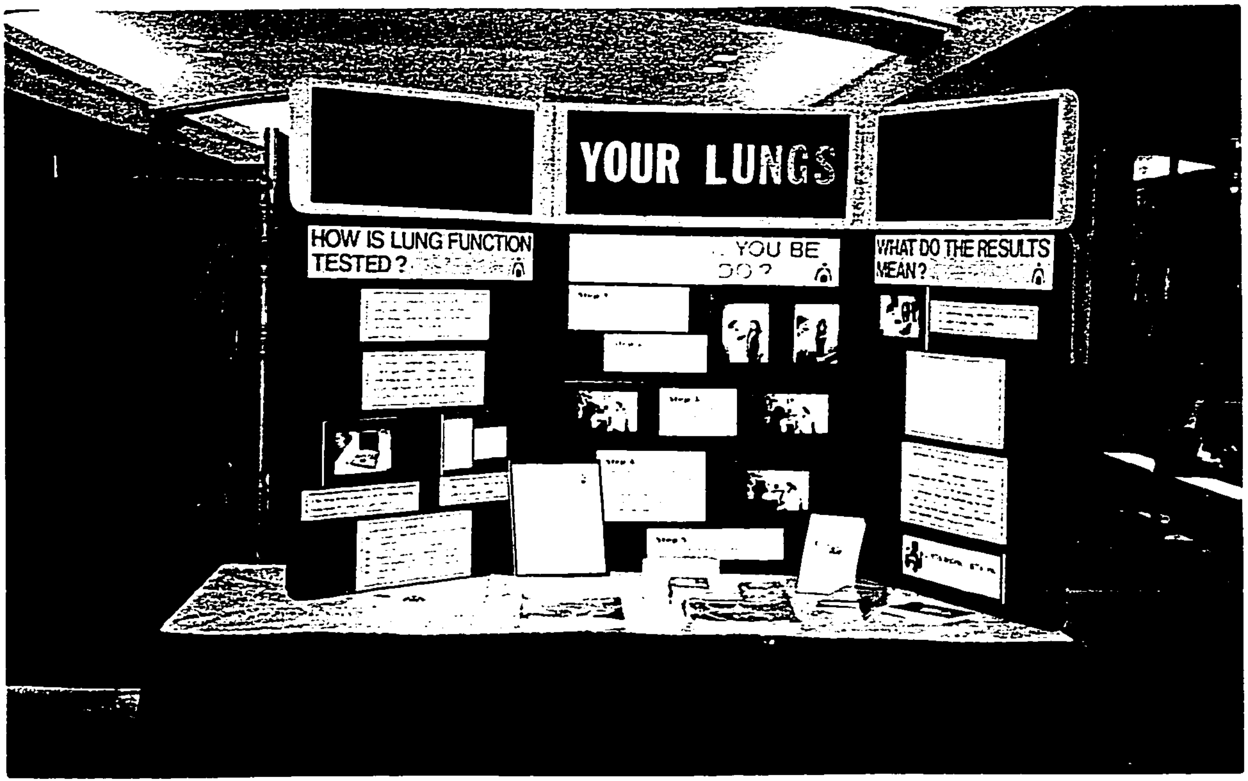


Figure 2a4. Study Display outlining spirometric testing procedures

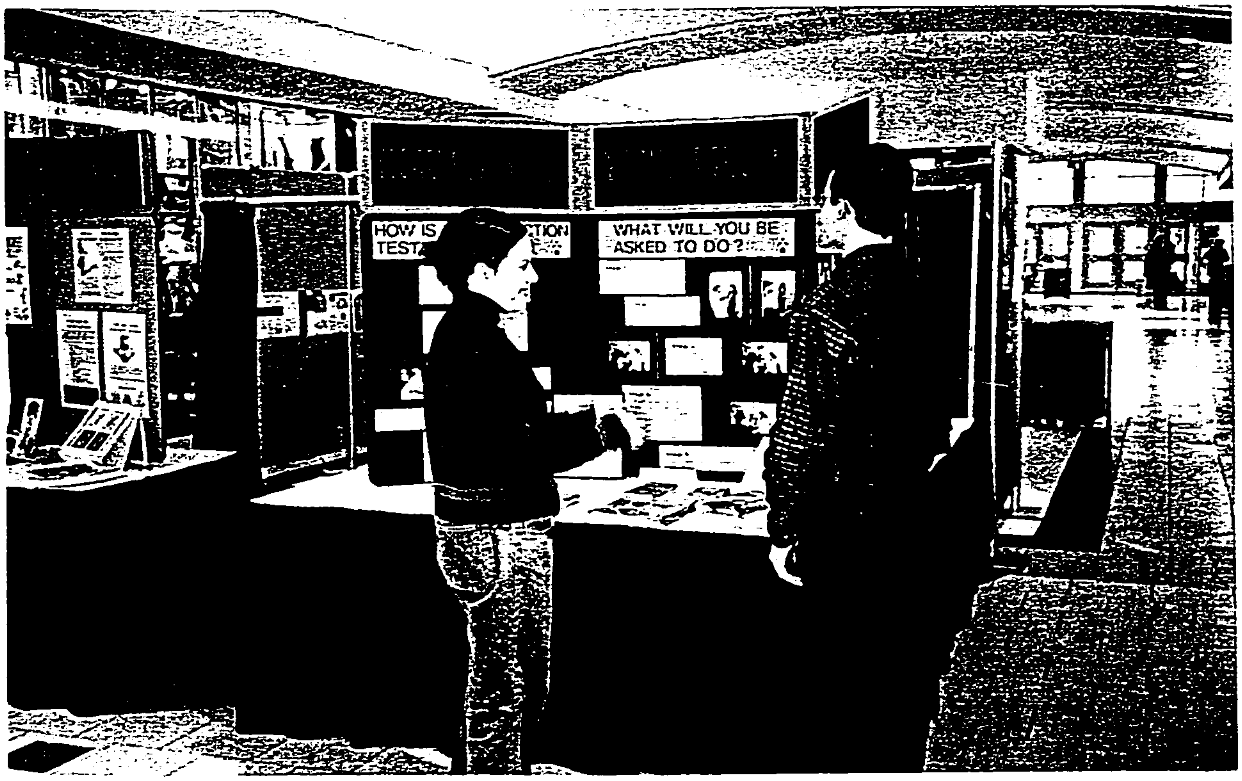


Figure 2a5. Potential participant receiving information regarding the study

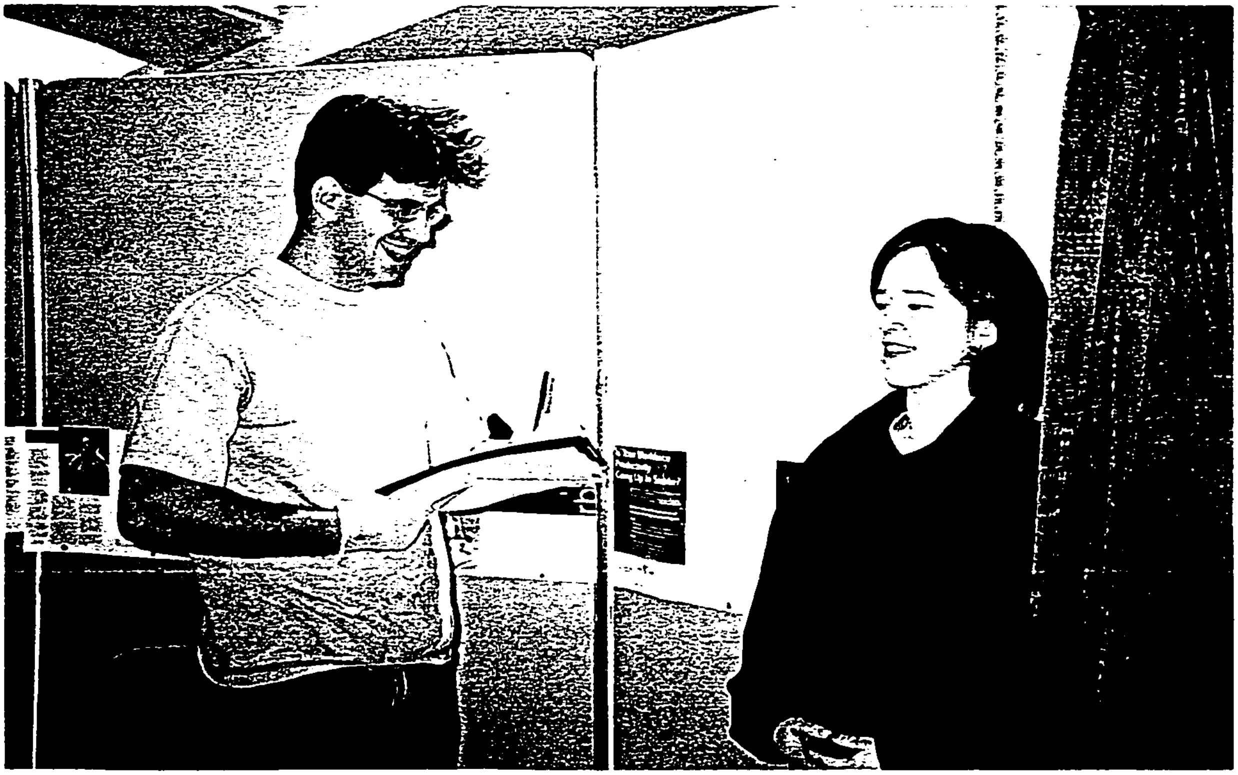


Figure 2a6. Study participant completing the technician-administered questionnaire

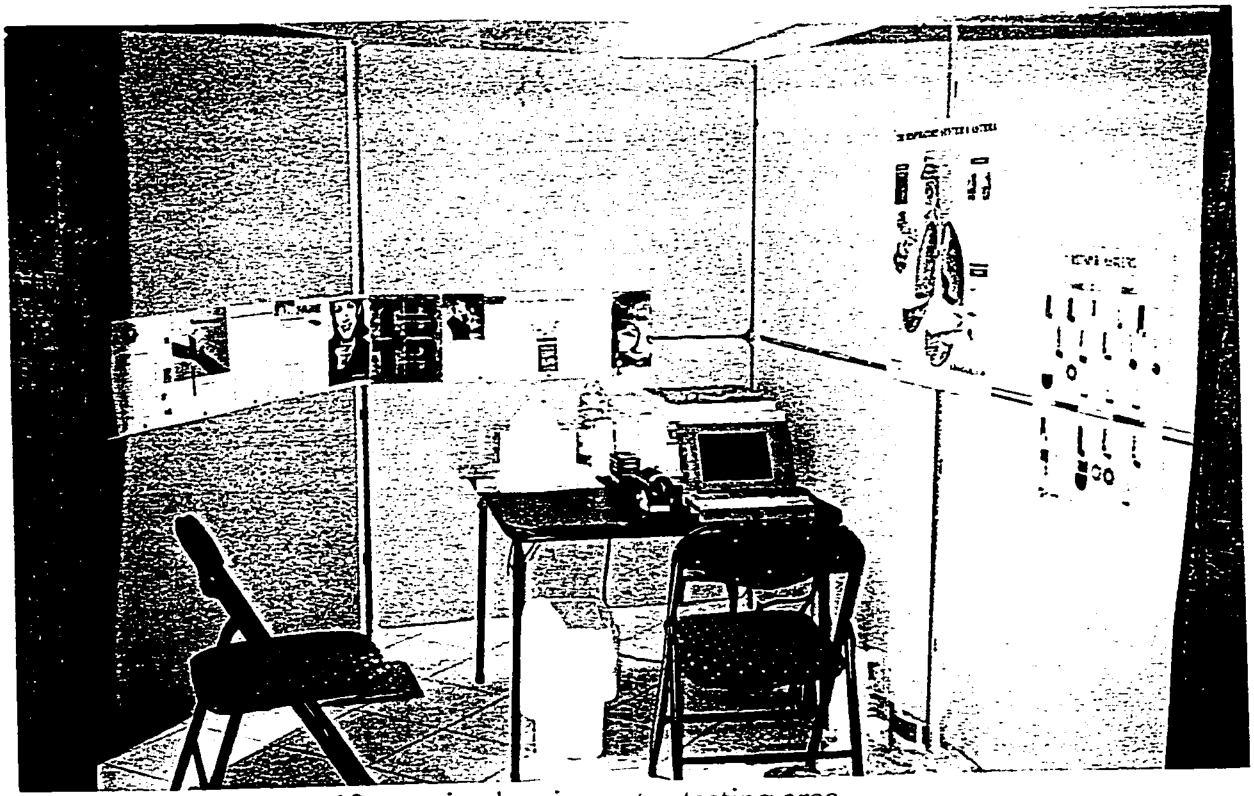


Figure 2a7. Private, self-contained, spirometry testing area

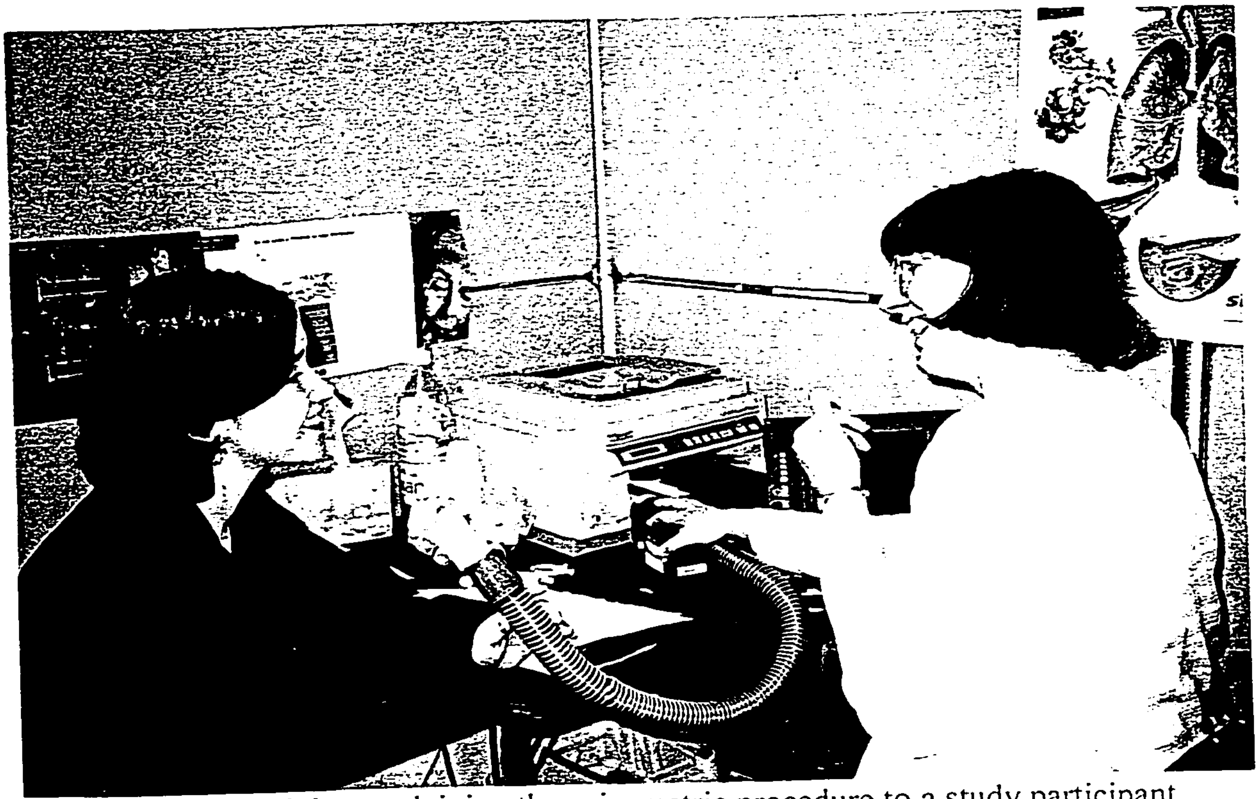


Figure 2a8. Technician explaining the spirometric procedure to a study participant

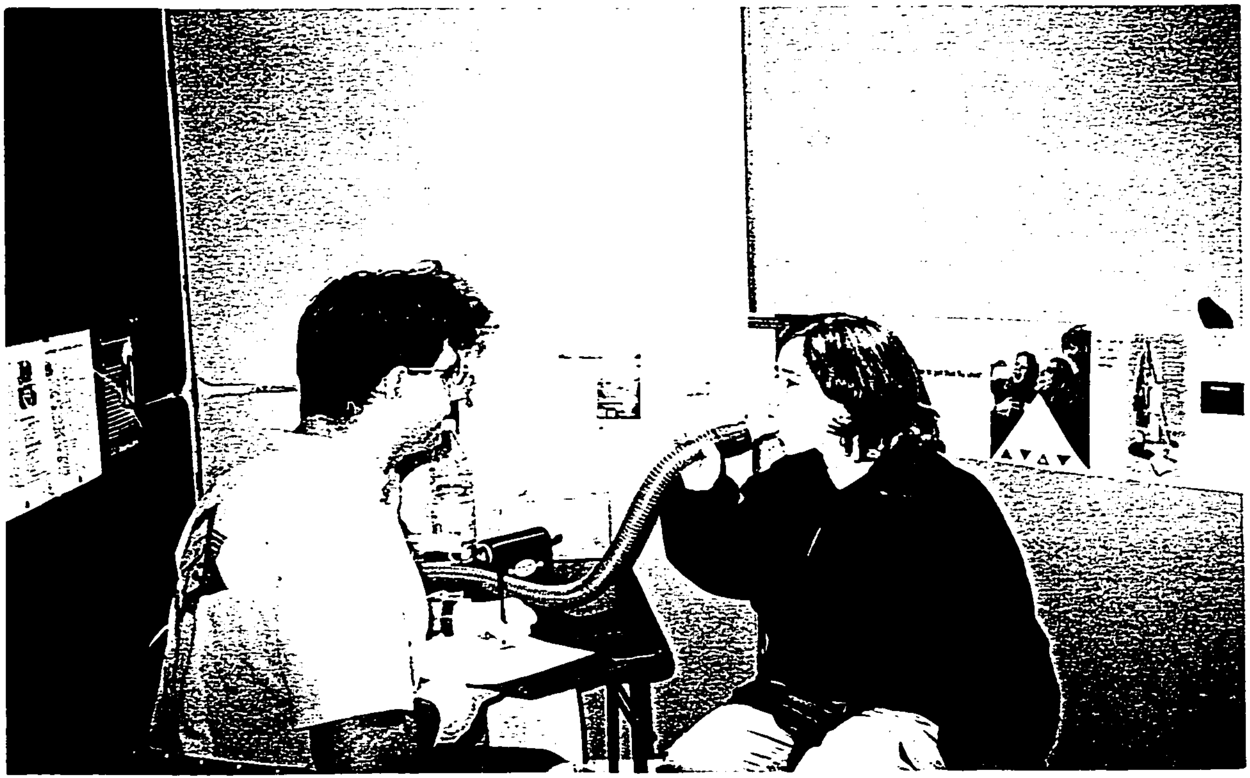


Figure 2a9. Technician coaching a participant during the testing procedure

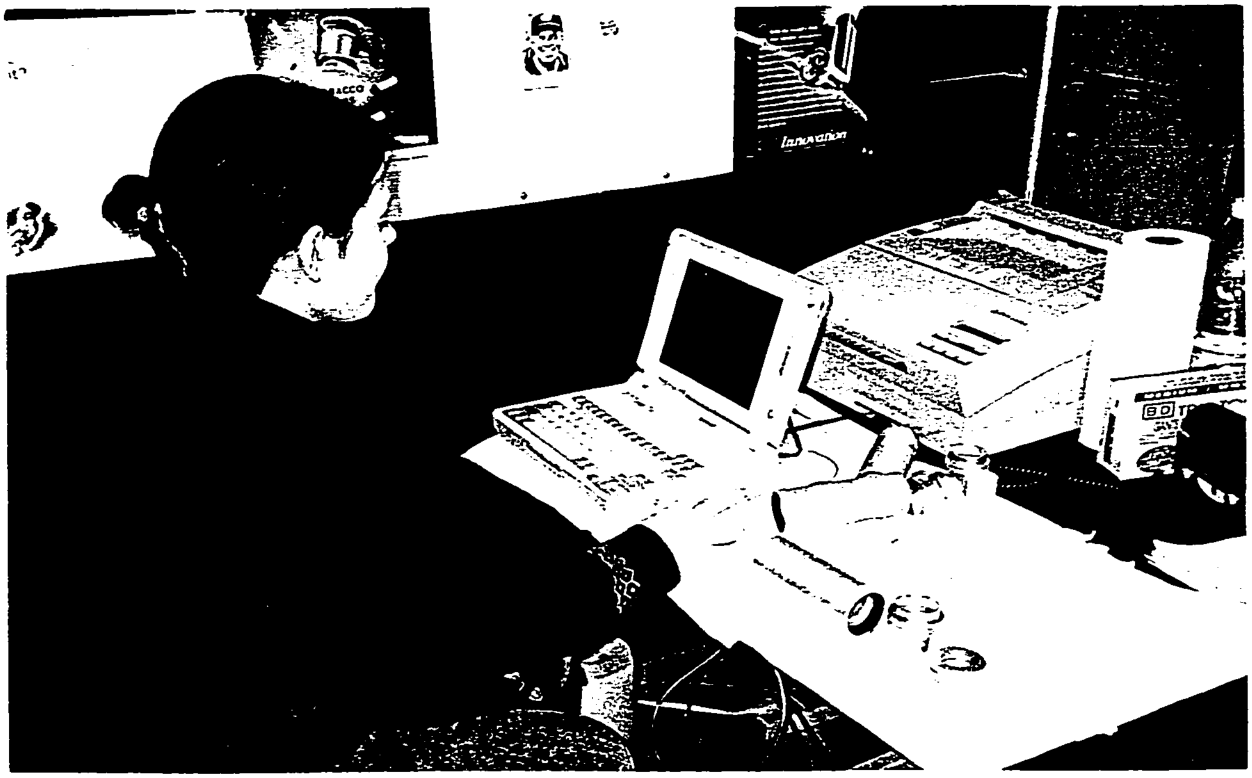


Figure 2a10. Retrieving results for analysis

Appendix 3:

Participant Information Letter

Primary Data Set

INFORMATION LETTER

TITLE OF PROJECT:	**Impact of Raters' Levels of Training and Experience on the Quality of Spirometric Interpretations in Epidemiological Studies**	
PRINCIPAL INVESTIGATOR:		
Dr. Patrick Hessel	**Department of Public Health Sciences University of Alberta**	**407-7135**
CO-INVESTIGATORS:		
Dr. Richard L. Jones	**Department of Pulmonary Medicine University of Alberta**	**492-6475**
Dr. Don Schopflocher	**Alberta Health**	**422-4630**
Dr. Gus Thompson	**Department of Public Health Sciences University of Alberta**	**492-8753**
GRADUATE STUDENT:		
Tania Stafinski	**Department of Public Health Sciences University of Alberta**	**407-6654**

February 5, 1999

Dear Participant:

You are being asked to take part in a study that will look at the repeatability of the interpretation of lung function tests.

The lung function test is called spirometry. Spirometry measures how much air you can blow out of your lungs and how fast you can blow that air out after a full inhalation. The record of the test is called a spirogram. These results are often used to detect lung problems.

Before the spirogram can be used, it must be decided if the test was performed correctly. This is done by looking at standard guidelines that have been developed. If the test does not meet these guidelines, it is rejected and the spirogram cannot be used. Even with these guidelines, judgement of a spirogram's acceptability can still vary. Therefore, the amount of agreement between people evaluating the spirograms is important when trying to decide if the results are reliable.

Purpose of the study

The purpose of this study is to determine the level of training and experience needed to correctly decide whether or not spirograms are acceptable. This information will help to make sure that studies using spirometry in the future will be of high quality

The data are also being collected as part of a graduate thesis project.

Background:
Someone who is trained in directing lung function tests will show you what to do. The spirograms will then be given to several interpreters who will each look at them and decide if they are acceptable.

Participating in the study will involve:
Two lung function technicians will carry out the spirometric test. It will take about 20 minutes to complete. You will be tested by one technician first. and then by the other.

The technician will first show you how the test is done. He or she will make sure you understand what you are being asked to do.

The technician will measure your height and weight and ask you your age. gender. smoking status. and ethnic background. This information will be recorded.

The technician will ask you to take a deep breath in and to blow into the spirometer following his or her instructions.

Testing will take place at a local shopping mall during the middle of February,1999.

Confidentiality and voluntary participation:
All records will be kept private. Only research investigators will have access to your spirograms. Before your results are given to the interpreters to evaluate, your name will be removed. Consent forms and interpretation forms will be kept in a secure area for at least seven years. You are free to withdraw from this study at any time.

Possible benefits to you and others:
At the end of the test, the technician will discuss your results with you. The technician will tell you how your results compare to those of other people your age. In addition. if they appear to be abnormal, the technician will inform you and refer you to an appropriate health care professional for follow-up.

By determining the type of training and experience needed to correctly judge spirograms for acceptability, this study will help to set standards for data collected in future lung function studies.

Possible risks:
There are no expected risks by taking part in this study.

For further information on the study:
If you have any questions about this study or would like to have further details during the study, please contact: Patrick Hessel at 407-7135 or Dennis Michaelchuk, Alberta Asthma Centre (not involved in study) at 407-7097.

Your consent:
Your signature on the next page indicates that you understand the information about participation in this study and also that you agree to be involved.

I have read the above information letter:

________________________Initials of study participant

________________________Initials of researcher

Appendix 4:

Participant Consent Form

Primary Data Set

PARTICIPANT CONSENT FORM

TITLE OF PROJECT: **Impact of Raters' Levels of Training and Experience on the Quality of Spirometric Interpretations in Epidemiological Studies**

PRINCIPAL INVESTIGATOR:		
Dr. Patrick Hessel:	**Department of Public Health Sciences** **University of Alberta**	**407-7135**
CO-INVESTIGATORS:		
Dr. Richard L. Jones	**Department of Pulmonary Medicine** **University of Alberta**	**492-6475**
Dr. Don Schopflocher	**Alberta Health**	**422-4630**
Dr. Gus Thompson	**Department of Public Health Sciences** **University of Alberta**	**492-8753**
GRADUATE STUDENT:		
Tania Stafinski	**Department of Public Health Sciences** **University of Alberta**	**407-6654**

Do you understand that you have been asked to be in a research study?	□ Yes	□ No
Have you read and received a copy of the attached Information Sheet?	□ Yes	□ No
Do you understand the benefits and risks involved in taking part in this research study?	□ Yes	□ No
Have you had an opportunity to ask questions and discuss this study?	□ Yes	.□No
Do you understand that you are free to refuse to participate or withdraw from the study at any time? You do not have to give a reason.	□ Yes	□ No
Has the issue of confidentiality been explained to you?	□ Yes	□ No
Do you understand who will have access to your information?	□ Yes	□ No
I agree to take part in this study.	□ Yes	□ No

Signature of Research Participant:______________________________ Date:____________

Phone: ______________(Work)

______________ (Home)

I believe that the person signing this form understands what the study involves and voluntarily agrees to participate.

Signature of Investigator or Designee:__________________________ Date:____________

Appendix 5:

Lung Function Testing Questionnaire

Primary Data Set

Participant's Identification Number:_______

Lung Function Testing Questionnaire

Participant's Name: ______________________________

Participant's Phone Number: ______________________

Participant's Date of Birth: ______________________
(day/month/year)

Participant's Gender: □ Male □ Female

Spirometry Questions:

		Yes	No
1.	Have you ever done spirometry?	□	□

If yes, go to question 2.
If no, go to Smoking Status Questions.

2. When was the last time you did spirometry? ________________

3. How many times have you done spirometry? ________________

Smoking Status Questions:

		Yes	No
1.	Have you ever smoked for as long as a year?	□	□

(This means at least one cigarette per day or one cigar per week for one year.)

If yes, go to question 2.
If no, stop here.

2. How old were you when you started smoking? ________________
(years)

		Yes	No
3.	Do you now smoke, as of one month ago?	□	□

If yes, go to question 6.
If no, **only** answer questions 4 and 5.

4. How old were you when you quit smoking? ________________
(years)

5. On average, of the entire time you smoked, how much did you smoke?

	Number
A. Cigarettes per day	______
B. Cigarillos per day	______
C. Cigars per day	______
D. Pipe tobacco per week (oz)	______
E. Pipe tobacco per week (gm)	______

	Yes	No
6. Have you cut down smoking?	□	□

If yes, **only** answer questions 7 and 8.
If no, go to question 9.

7. How old were you when you cut down smoking? ______________
(years)

8. On average, of the entire time you cut down, how much did you smoke?

	Number
A. Cigarettes per day	______
B. Cigarillos per day	______
C. Cigars per week	______
D. Pipe tobacco per week (oz)	______
E. Pipe tobacco per week (gm)	______

9. On average, how much do you smoke?

	Number
A. Cigarettes per day	______
B. Cigarillos per day	______
C. Cigars per week	______
D. Pipe tobacco per week (oz)	______
E. Pipe tobacco per week (gm)	______

Appendix 6:

Spirometry Instruction for the Minimally Trained Technician

Spirometry Instruction Sessions

Time Required: 10 hours (2 - 5 hour days)

Day 1

Hours One and Two

Introduction to Spirometry

1. Overview of spirometry
   - A. Definition of spirometry
   - B. Indications for spirometry
   - C. Discussion of variables to be measured:
     - i. FEV_1, FVC, FEF_{25-75}, PEFR
     - ii. Labeling of spirograms, flow-volume curves, and volume-time curves
     - iii. Physiology of flow-volume loops
   - D. Brief explanation of derivation of percent of predicted values
   - E. Interpretation of Results
     - i. Characteristics of curves indicative of obstructive or restrictive defects
2. Quality of Spirometric Results
   - A. Importance of performing spirometry properly
   - B. Presentation and application of ATS criteria
     - i. Definition of acceptability and reproducibility
     - ii. Examples of satisfactory vs unsatisfactory expiratory flow curves

Hours Three and Four

Introduction to Spirometry Procedures

1. Operation of spirometry equipment
   - A. Set-up of device
   - B. Calibration instructions
     - i. Completion of calibration test using a calibration syringe

2. Performance of spirometry
    A. Entry of patient data
        i. Collection of demographic information and cardiopulmonary history
    B. Explanation and demonstration of how to complete the spirometric maneuver
    C. Storage of Results
3. Demonstration of the entire testing procedure

Hour Five

Practice procedures using volunteer participants

Day 2

Hour Six

Review of instructions for performing spirometric maneuvers

1. Review of spirometry terms
2. Discussion of ATS acceptability and reproducibility standards
    A. Overview of "start of test" criteria
    B. Overview of "end of test" criteria

Hour Seven

Trouble-shooting techniques

1. Recognition of "problem" spirograms
2. Importance of coaching to obtain maximum effort from participants
3. Presentation of case scenarios
    A.. Strategies for obtaining acceptable spirometric results from different types of patients
    B. Corrective techniques

Hours Eight and Nine

Practice administering spirometry to volunteer participants

Hour Ten

Setting up a Field Clinic

1. Review of equipment required
2. Maintenance of equipment
   - A. Proper cleaning protocol
   - B. Identification and correction of common instrument problems

Appendix 7:

Participant Pulmonary and Cardiovascular Health History

Primary Data Set

Cardiopulmonary Health Questionnaire

	Yes	No
1. Is the patient ever short of breath?	□	□
When at rest:	□	□
While walking:	□	□
On stairs:	□	□
2. Does the patient experience wheezing or asthma?	□	□
3. Does the patient have a history of heart disease?	□	□
4. Has the patient had an abnormal chest X-Ray?	□	□
5. Does the patient experience pain while breathing?	□	□
6. Has the patient had lung surgery?	□	□
What kind?		
One lung removed	□	□
Lobectomy	□	□
Biopsy	□	□
Other ______________	□	□
7. Does the patient cough frequently?	□	□
With sputum?	□	□

Appendix 8:

Participant Introduction Letter

Participant Information and Consent Forms

Secondary Data Set

INFORMATION LETTER

TITLE OF PROJECT: **Impact of Raters' Levels of Training and Experience on the Quality of Spirometric Interpretations in Epidemiological Studies**

PRINCIPAL INVESTIGATOR:		
Dr. Patrick Hessel	**Department of Public Health Sciences** **University of Alberta**	**407-7135**
CO-INVESTIGATORS:		
Dr. Richard L. Jones	**Department of Pulmonary Medicine** **University of Alberta**	**492-6475**
Dr. Don Schopflocher	**Alberta Health**	**422-4630**
Dr. Gus Thompson	**Department of Public Health Sciences** **University of Alberta**	**492-8753**
GRADUATE STUDENT:		
Tania Stafinski	**Department of Public Health Sciences** **University of Alberta**	**407-6654**

November 11, 1998

Dear Faculty, Staff, and Graduate Students:

You are being asked to take part in a study that will look at the repeatability of the interpretation of lung function tests.

The lung function test is called spirometry. Spirometry measures how much air you can blow out of your lungs and how fast you can blow that air out after a full inhalation. The record of the test is called a spirogram. These results are often used to detect lung problems.

Before the spirogram can be used, it must be decided if the test was performed correctly. This is done by looking at standard guidelines that have been developed. If the test does not meet these guidelines, it is rejected and the spirogram cannot be used. Even with these guidelines, judgement of a spirogram's acceptability can still vary. Therefore, the amount of agreement between people evaluating the spirograms is important when trying to decide if the results are reliable.

Purpose of the study

The purpose of this study is to determine the level of training and experience needed to correctly decide whether or not spirograms are acceptable. This information will help to make sure that studies using spirometry in the future will be of high quality.

The data are also being collected as part of a graduate thesis project.

Background:
Fifty graduate students and 50 employees from Alberta Health will participate in the study by doing spirometry. Someone who is trained in conducting lung function tests will show you what to do. The spirograms will then be given to several interpreters who will each look at them and decide if they are acceptable.

Participating in the study will involve:
A certified pulmonary function technician will carry out the spirometric test. It will take about 15 minutes to complete.

The technician will first show you how the test is done. He or she will make sure you understand what you are being asked to do.

The technician will ask your age, gender, and ethnic background, and will measure your height and weight. This information will be recorded.

The technician will ask you to take a deep breath in and to blow into the spirometer following his or her instructions.

Testing will take place from November 23rd through November 27th from 2:00 p.m. to 5:00 p.m. in Patrick Hessel's office in the Department of Public Health Sciences. You will be contacted and asked to select one 15 minute time slot from this schedule that is most convenient for you. Please complete the attached consent form and place it in Tania Stafinski's mailbox in 13-109 as soon as possible.

Confidentiality and voluntary participation:
All records will be kept private. Only research investigators will have access to your spirograms. Before your results are given to the interpreters to evaluate, your name will be removed. Consent forms and interpretation forms will be kept in a secure area for at least seven years. You are free to withdraw from this study at any time.

Possible benefits to you and others:
At the end of the test, the technician will discuss your results with you. The technician will tell you how your results compare to those of other people your age. In addition, if they appear to be abnormal, the technician will inform you and refer you to an appropriate health care professional for follow-up.

By determining the type of training and experience needed to correctly judge spirograms for acceptability, this study will help to set standards for data collected in future lung function studies.

Possible risks:
There are no expected risks by taking part in this study.

For further information on the study:
If you have any questions about this study or would like to have further details during the study, please contact: Patrick Hessel at 407-7135 or Dennis Michaelchuk, Alberta Asthma Centre (not involved in the study) at 407-7097.

Your consent and legal rights:
Your signature on the next page indicates that you understand the information about participation in this study and also that you agree to be involved.

I have read the above information letter:

_________Initials of study participant

_________Initials of researcher

PARTICIPANT CONSENT FORM

TITLE OF PROJECT:	Impact of Raters' Levels of Training and Experience on the Quality of Spirometric Interpretations in Epidemiological Studies	
PRINCIPAL INVESTIGATOR:		
Dr. Patrick Hessel:	Department of Public Health Sciences University of Alberta	407-7135
CO-INVESTIGATORS:		
Dr. Richard L. Jones	Department of Pulmonary Medicine University of Alberta	492-6475
Dr. Don Schopflocher	Alberta Health	422-4630
Dr. Gus Thompson	Department of Public Health Sciences University of Alberta	492-8753
GRADUATE STUDENT:		
Tania Stafinski	Department of Public Health Sciences University of Alberta	407-6654

Do you understand that you have been asked to be in a research study? □ Yes □ No

Have you read and received a copy of the attached Information Sheet? □ Yes □ No

Do you understand the benefits and risks involved in taking part in this research study? □ Yes □ No

Have you had an opportunity to ask questions and discuss this study? □ Yes □No

Do you understand that you are free to refuse to participate or withdraw from the study at any time? You do not have to give a reason. □ Yes □ No

Has the issue of confidentiality been explained to you? □ Yes □ No

Do you understand who will have access to your information? □ Yes □ No

I agree to take part in this study. □ Yes □ No

Signature of Research Participant:____________________________ Date:____________

Phone: ______________(Work)

______________ (Home)

I believe that the person signing this form understands what the study involves and voluntarily agrees to participate.

Signature of Investigator or Designee:_________________________ Date:____________

Appendix 9:

Participant Health Assessment Questionnaire

Secondary Data Set

Participant's Identification Number:______

Lung Function Testing Questionnaire

Participant's Name:______________________________

Participant's Phone Number:______________________

Participant's Date of Birth:_______________________
(day/month/year)

Participant's Gender: □ Male □ Female

Smoking Status Questions:

Yes No

1. Have you ever smoked for as long as one year? □ □

(This means at least one cigarette per day or one cigar per week for one year.)

If yes, go to question 2.
If no, stop here.

2. How old were you when you started smoking? ____________
(years)

Yes No

3. Do you now smoke, as of one month ago? □ □

If yes, go to question 6.
If no, **only** answer questions 4 and 5.

4. How old were you when you quit smoking? ___________
(years)

5. On average, of the entire time you smoked, how much did you smoke?

	Number
A. Cigarettes per day	_________
B. Cigarillos per day	_________
C. Cigars per week	_________
D. Pipe tobacco per week (oz)	_________
E. Pipe tobacco per week (gm)	_________

(Lung Function Questionnaire continued)

	Yes	No
6. Have you cut down smoking?	□	□

If yes, **only** answer questions 7 and 8.
If no, go to question 9.

7. How old were you when you cut down smoking? __________
(years)

8. On average, of the entire time you cut down, how much did you smoke?

Number
A. Cigarettes per day __________
B. Cigarillos per day __________
C. Cigars per week __________
D. Pipe tobacco per week (oz) __________
E. Pipe tobacco per week (gm) __________

9. On average, how much do you smoke?

Number
A. Cigarettes per day __________
B. Cigarillos per day __________
C. Cigars per week __________
D. Pipe tobacco per week (oz) __________
E. Pipe tobacco per week (gm) __________

Appendix 10:

Spirogram Interpretation Forms

Spirogram
Identification Number: ______

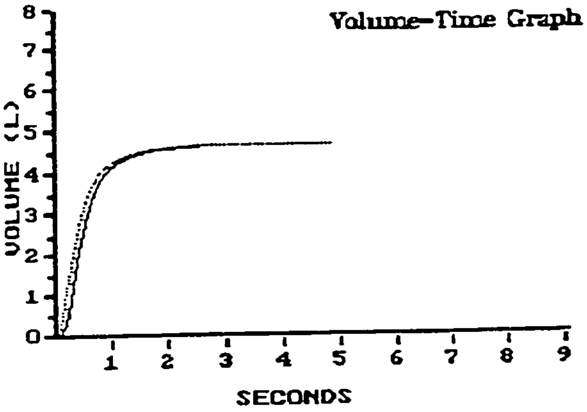
Interpreter
Identification Number: ______

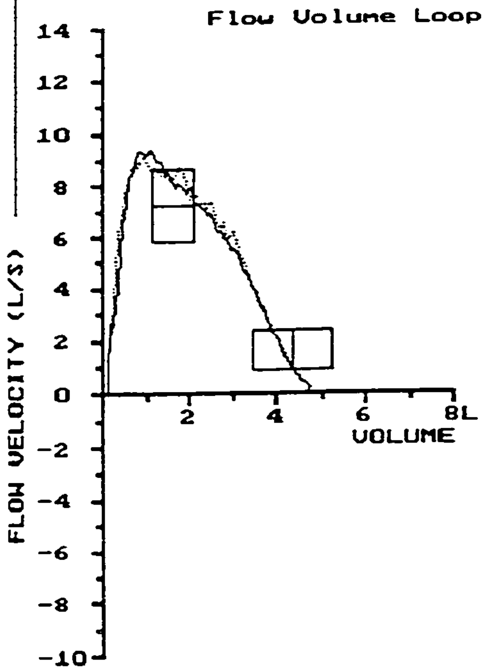
Spirogram Interpretation Form

Please examine the Flow-Volume Loop, Volume-Time Curve, and lung function values below:

Expiratory	Actual	Predicted	% of pred
FVC	4.65 L	4.33 L	107.39 %
FEV 0.5	3.47 L	2.82 L	122.74 %
FEV 1.0	4.34 L	3.67 L	118.19 %
FEV 3.0	4.65 L	3.84 L	120.96 %
FEV 0.5/FVC	74.58 %	65.25 %	114.30 %
FEV 1.0/FVC	93.30 %	84.77 %	110.06 %
FEV 3.0/FVC	100.00 %	88.78 %	112.64 %
PEF	9.32 L/S	7.04 L/S	132.32 %
FEF 25-75%	6.28 L/S	3.84 L/S	163.32 %
FEF 75-85%	2.74 L/S	1.48 L/S	185.47 %
FEF 25	8.89 L/S	6.34 L/S	140.16 %
FEF 50	6.78 L/S	4.49 L/S	151.02 %
FEF 75	3.72 L/S	2.10 L/S	176.61 %
FEF .2-1.2	7.58 L/S	6.39 L/S	118.59 %

Inspiratory	Actual
IVC	0.00 L
FIV1	0.00 L
PIF	0.00 L/S
FIF50	0.00 L/S
FEF50/FIF50	0.00 %





	Yes	No
Overall, based on the above graphic results, was the spirometric test acceptable?	□	□

If you found the test not to be acceptable, please briefly state your reasons below:

If you found the test to be acceptable, but questioned one or more features of the the spirograms, please briefly state them below:

Appendix 11:

Rater Information Letter

Information Letter

Impact of Raters' Levels of Training and Experience on the Quality of Spirometric Interpretations in Epidemiologic Studies

PRINCIPAL INVESTIGATOR:		
Dr. Patrick Hessel	**Department of Public Health Sciences** **University of Alberta**	**407-7135**
CO-INVESTIGATORS:		
Dr. Richard L. Jones	**Department of Pulmonary Medicine** **University of Alberta**	**492-6475**
Don Schopflocher	**Alberta Health**	**422-4630**
Dr. Gus Thompson	**Department of Public Health Sciences** **University of Alberta**	**492-8753**
GRADUATE STUDENT:		
Tania Stafinski	**Department of Public Health Sciences** **University of Alberta**	**407-6654**

May 25, 1999

Dear Rater:

You are being asked to take part in a study that will examine the repeatability of the interpretation of spirometric results.

The validity of spirometric results depends upon correct interpretation of the spirograms. The first step in the interpretation process involves evaluating the spirogram's acceptability using criteria developed by the American Thoracic Society. Even with these guidelines, judgement of a spirogram's acceptability can vary. Therefore, the amount of agreement between people evaluating the spirograms is important when attempting to determine if the results are reliable

Purpose of the study

The purpose of this study is to determine the level of training and experience needed to correctly decide whether or not spirograms are acceptable. This information will help to ensure that studies using spirometry in the future will be of high quality.

The data are also being collected as part of a graduate thesis project.

Background

Four categories of interpreters will be assembled based upon their level of pulmonary expertise. Each rater will be asked to judge the acceptability/nonacceptability of a common set of spirograms. The level of agreement both within and between categories of interpreters will then be assessed.

Participating in the study will involve:

You will receive a set of spirograms from 400 subjects (two from each subject).

Following the instructions found on each spirogram interpretation form, you will be asked to complete all 800 sheets and return them in the package provided at your earliest convenience.

Confidentiality and voluntary participation

You will be assigned a confidential rater identification number. Your name will not appear on the interpretation form. All records of your interpretations will be kept private. Only research investigators will have access them. Further, both consent forms and interpretation forms will be stored in a secure area for at least seven years. You are free to withdraw from this study at any time.

Possible benefits to you and others

Study results will be forwarded to you as soon as they are available.

By determining the type of training and experience needed to correctly judge spirograms for acceptability, this study will assist in establishing appropriate standards for data collected in future research involving spirometry.

Possible risks

There are no expected risks in taking part in this study.

For further information on the study

If you have any questions or concerns about any aspect of this study, please contact:
Patrick Hessel at 492-4159 or Dennis Michaelchuk, Alberta Asthma Centre (not involved in the study) at 492-7097.

Your consent

Your signature on the next page indicates that you understand the information about participation in this study and also that you agree to be involved.

Please keep these pages for future reference.

I have read the above information letter:

______Initials of study participant (rater)

Appendix 12:

Instructional Letter Regarding Completion of Interpretation Forms

25-May-1999

Dear Interpreter,

Thank you for participating in our study entitled: “Impact of Raters’ Levels of Training and Experience on the Quality of Spirometric Interpretations in Epidemiologic Studies.”

Enclosed are 400 pairs of spirograms. The spirograms for each subject are included sequentially (i.e., spirograms 1 and 2 are from the same person, 3 and 4 are from the next person, etc.). For the most part, each spirogram should be evaluated independently (i.e., we are not asking you to assess reproducibility). The reason for juxtaposing the spirograms for the same individual is to facilitate interpretation of the acceptability of the individual tracings.

For each spirogram interpretation form, please complete the evaluation section. If the spirogram is clearly acceptable, simply place a tick mark in the corresponding box. If it is unacceptable, please indicate why (briefly) in the space provided on the form. If you ultimately decide that the spirogram is acceptable, but questioned its acceptability on the basis of one or more features, please note those feature(s) in the appropriate area on the form.

Please return the completed forms via courier collect mailing.

Feel free to contact me if you have any questions. Thank you very much for your time and assistance.

Very truly yours,

Patrick A. Hessel
Associate Professor
Director, Epidemiology Program

Appendix 13:

Rater Consent Form

SPIROGRAM RATER CONSENT FORM

TITLE OF PROJECT: **Impact of Raters' Levels of Training and Experience on the Quality of Spirometric Interpretations in Epidemiological Studies**

PRINCIPAL INVESTIGATOR:		
Dr. Patrick Hessel:	**Department of Public Health Sciences** **University of Alberta**	**407-7135**
CO-INVESTIGATORS:		
Dr. Richard L. Jones	**Department of Pulmonary Medicine** **University of Alberta**	**492-6475**
Dr. Don Schopflocher	**Alberta Health**	**422-4630**
Dr. Gus Thompson	**Department of Public Health Sciences** **University of Alberta**	**492-8753**
GRADUATE STUDENT:		
Tania Stafinski	**Department of Public Health Sciences** **University of Alberta**	**407-6654**

Do you understand that you have been asked to be in a research study? □ Yes □ No

Have you read and received a copy of the attached Information Sheet? □ Yes □ No

Do you understand the benefits and risks involved in taking part in this research study? □ Yes □ No

Have you had an opportunity to ask questions and discuss this study? □ Yes □No

Do you understand that you are free to refuse to participate or withdraw from the study at any time? You do not have to give a reason. □ Yes □ No

Has the issue of confidentiality been explained to you? □ Yes □ No

Do you understand who will have access to your information? □ Yes □ No

I agree to take part in this study. □ Yes □ No

Signature of Research Participant:______________________________ Date:_____________

Phone: ______________(Work)

______________ (Home)

I believe that the person signing this form understands what the study involves and voluntarily agrees to participate.

Signature of Investigator or Designee:__________________________ Date:_____________

Appendix 14:

American Thoracic Society's Standardization of Spirometry

1994

American Thoracic Society Standardization of Spirometry 1994 Update
Edited Excerpts from Recommendation: VC and FVC - Maneuver Acceptability

Acceptability Criteria:

1. Satisfactory start-of-test

   An unsatisfactory start of expiration is characterized by the observation of one or more of the following:

   a. Excessive hesitation
   b. False start
   c. Extrapolated volume of greater than 5% of FVC or 0.15 L, whichever is greater

2. Satisfactory end-of-test

   Early termination of expiration is characterized by:

   a. Absence of a plateau, as defined by no change in volume for at least one second or a reasonable expiratory time, in the volume-time curve
   Note: In a normal young subject this would be before completion of the breath - usualiy less than a 6 second maneuver
   In an obstructed or older healthy subject , a longer expiratory time is required to reach a plateau
   Multiple prolonged exhalations are seldom justified

3. No glottis closure or hesitation during the maneuver that causes a cessation of airflow

4. No leak

5. No Obstructed mouthpiece

   The technician should also observe that the subject understood the instructions and performed the maneuver with all of the following:

   a. Maximum inspiration
   b. A good start
   c. Smooth continuous exhalation
   d. Maximal effort

Appendix 15:

Collapsing of Codes into Meaningful Categories for Statistical Analysis

Codes used for analysis of responses to open-ended questions regarding the acceptability of curves on spirogram interpretation forms

Category	Code number	Code description
Start of test		
	01	Poor start of test
	02	Slow start
	03	Hesitating start
	04	Inhalation at start
	05	Questionable start time
	06	Questionable peak flow / FEFmax
	07	Questionable FEV_1 values
	08	Flow fluctuations within first second
	09	Greater than 5% volume of extrapolation
End of test		
	10	Poor termination - general
	11	Incomplete exhalation
	12	Questionable exhalation time
	13	Questionable termination
	14	No one second plateau
	15	Stable plateau reached
	16	Forced exhalation at end of test
Effort		
	20	Variable effort - general
	21	Variable effort within first second of test
	22	Variable effort not affecting results
Cough		
	30	Cough - general
	31	Cough at start of test
	32	Cough at end of test

Codes used for analysis of responses to open-ended questions regarding the acceptability of curves on spirogram interpretation forms cont'd

Category	Code number	Code description
Glottis closure		
	40	Glottis closure
	41	Cough or glottis closure
Leak		
	50	Leak or loss of seal
Obstruction		
	60	Tongue obstruction / biting on mouthpiece
	61	Flow obstruction
	62	Airway obstruction
Other		
	70	Extra inhalation during test
	71	Dynamic compression
	72	More maneuvers required
General Artifacts of Curves		
	80	Flow fluctuations - general
	81	Artifact on tracing
	82	Questionable flow-volume loop
	83	Questionable volume-time loop
	84	Sharp increase in volume near start of test
	85	Sharp decrease in volume near end of test
Technical		
	90	Equipment problem
	91	Noise
	92	Inconsistent data
	93	Recalculations required

Appendix 16:

Selection of a Summary Measure for Kappa

Selection of a Method for Summarizing Kappa Values Within and Between Categories

Rationale For Using the Arithmetic Mean:

In any study, collection of data that are not normally distributed precludes the use of the arithmetic mean as an accurate measure of central tendency (Daniel, 1995). Therefore, the appropriateness of calculating average kappa coefficients was assessed by comparing values to those obtained through an "average of ranks"approach (which does not assume that the data are normally distributed). The average of ranks was computed as follows:

1. Individual kappa scores (displayed on Tables 4.1 to 4.4 and 4.7 to 4.14) were ranked according to their magnitude (from smallest to largest).
2. Ranks for sets of pairs representing within or between category agreement were averaged.

These values, in addition to arithmetic mean scores, were separately placed in ascending order and labeled 1 through 10. Table 16.1 displays results for kappa values presented in Table 4.3. Average rank values and average kappa coefficients, when matched on order, corresponded to identical pairs of raters. Consequently, the use of average kappa scores was deemed appropriate.

Table 16.1 Level of Agreement Between Raters (Expressed As Cohen's Kappa) For the Acceptability of Participants' Best Test Spirograms in the Secondary Data Set

	Non-Certified, Minimally Trained, Respiratory Research Assistants (RA)	Certified Respiratory Technicians (CRT)	Respiratory Epidemiologists (RE)	Pulmonary Specialists (PS)
RA	Arithmetic Average: 0.557 Order: 1 Average of Ranks: 39.3 Order: 1	Arithmetic Average: 0.261 Order: 10 Average of Ranks: 15.8 Order: 10	Arithmetic Average: 0.358 Order 6 Average of Ranks: 22.6 Order: 6	Arithmetic Average: 0.318 Order: 8 Average of Ranks: 18.3 Order: 9
CRT	n/a	Arithmetic Average: 0.337 Order: 7 Average of Ranks: 20.8 Order: 7	Arithmetic Average: 0.302 Order: 9 Average of Ranks: 19.9 Order: 8	Arithmetic Average: 0.382 Order: 5 Average of Ranks: 26.9 Order: 5
RE	n/a	n/a	Arithmetic Average: 0.480 Order: 2 Average of Ranks: 36.5 Order: 2	Arithmetic Average: 0.448 Order: 3 Average of Ranks: 30.4 Order: 3
PS	n/a	n/a	n/a	Arithmetic Average: 0.400 Order: 4 Average of Ranks: 27.5 Order: 4

Appendix 17:

Logistic Regression Variables and Model Building Steps

Table 17.1 Codes of Variables Used in Logistic Regression Analyses

Variable Description	Variable Abbreviation	Coding Scheme
Variables In Common With Both Data Sets		
Dependent Variable (outcome):		
Acceptability of "Best Test" Spirogram	BESTACPT	0 - Not Acceptable 1 - Acceptable
Independent Variables (covariates):		
Gender of Participant	GENDER	0 - Male 1 - Female
Age - continuous - dichotomous	AGE50	0 - < 50 years 1 - ≥ 50 years
Smoking History in Pack-years - continuous - categorical	PACKCAT	0 - < 1 pack-year 1 - ≤ pack-years > 10 2 - ≥ 10 pack-years
Expertise of Spirometry Technician	TECH	0 - Non-certified, Minimally Trained 1 - Certified , Highly Trained
Percent of Predicted FEV_1 Lung Function Value from Best Test Spirogram	FEVPCTC	0 - < Lower Limit of Normal (80% of predicted) 1 - ≥ Lower Limit of Normal (80% of predicted)
Percent of Predicted FCV Lung Function Value From Best Test Spirogram	FVCPCTC	0 - < Lower Limit of Normal (80% of predicted) 1 - ≥ Lower Limit of Normal (80% of predicted)
Acceptability of Complementary Spirogram	ACPT2	0 - Not Acceptable 1 - Acceptable
ATS Reproducibility of Best Test Spirogram's Numeric Values	ATSREPRO	0 - Does Not Meet ATS Criteria 1 - Meets ATS Criteria
Variables Exclusive To the Primary Data Set		
History of Shortness of Breath	SBREATH	0 - No 1 - Yes
History of Painful Breathing	PBREATH	0 - No 1 - Yes
History of Wheeze or Asthma	WHEEZE	0 - No 1 - Yes
History of Frequent Cough	FRCOUGH	0 - No 1 - Yes
Abnormal Chest X-Ray	ABCHEST	0 - No 1 - Yes
Prior Completion of Spirometric Test	SPIROEXP	0 - No 1 - Yes

Unconditional Logistic Regression of the Primary Data Set: Model Building Steps

```
Method: Forward Stepwise (LR) ( Statistical Software: SPSS Version 9.0)
Step One:

 Total number of cases: 100 (unweighted)
 Number of cases included in the analysis: 100

 Initial -2 Log Likelihood: 128.207
 * Constant is included in the model

 ------------------------------------------------Variables in the Equation---------------------------------------------

                                                                                       95% CI for Exp(B)
 Variable      B       S.E.       Wald     df    Sig.       R          Exp(B)      Upper    Lower
 ---------------------------------------------------------------------------------------------------
 Constant   -0.663    0.211      9.870     1     0.002     NA          NA               NA

 --------------------------------------------Variables not in the Equation---------------------------------------------

 Residual Chi Square (3 df):  49.024     Sig. = 0.000

        Variable           Score       df      Sig

        AGE50               2.451       1     0.117

        ACPT2              46.477       1     0.000

       PACKCAT1             1.753       1     0.186

 Step Two:

 Variable Entered on Step Number 2: ACPT2
 - 2 Log Likelihood: 80.746
   Chi -Square (df 1): 48.990
   Significance: 0.000

 ---------------------------------------- -------Variables in the Equation---------------------------------------------

                                                                                        95% CI for Exp(B)
 Variable      B        S.E.       Wald     df     Sig.        R        Exp(B)      Upper     Lower
 ---------------------------------------------------------------------------------------------------
  ACPT2      3.523     0.606     33.774     1     0.000     0.4978     33.889     10.228 - 1111.192

 Constant   -1.914     0.357     28.721     1     0.000       NA         NA              NA

 ---------------------------------------------------Model if Term Removed----------------------------------------------

 Term Removed      Log Likelihood      -2 Log LR       df     Significance of Log LR

 ACPT2                -64.104            47.461        1             0.000

 --------------------------------------------Variables not in the Equation---------------------------------------------

 Residual Chi Square (2 df): 0.333      Sig. = 0.847

        Variable           Score       df      Sig

        AGE50              0.1788       1     0.6743

       PACKCAT1            0.1766       1     0.6724
No more variables can be added or deleted.
```

Exact Logistic Regression of the Primary Data Set: Model Building Steps

Method: Backward Elimination (Statistical Software: LogXact Version 2.1)

Step One:

Model: BESTACPT =AGE50+PACKCAT1+ACPT2
Stratum: <Unstratified>
Weight: <none>
Number of Observations: 100
Likelihood Ratio Statistic (4 df): 63.391

	INFERENCE	<-------------PARAMETER ESTIMATION------------>			P-VALUE
TERM	TYPE	ODDS RATIO	S.E.	95.0 % CONF. INTERVAL	2*1_SIDED
AGE50	Asymptotic	0.780	NA	0.2331 - 2.616	0.688
	Exact	0.800	NA	0.200 - 3.144	0.945
ACPT2	Asymptotic	31.814	NA	9.491 - 105.224	0.000
	Exact	26.755	NA	7.853 - 113.249	0.000
PACKCAT1	Asymptotic	0.729	NA	0.155 - 3.442	0.690
	Exact	0.742	NA	0.106 - 4.083	0.999
CONST	Asymptotic	0.174	NA	0.072 - 0.419	0.000

Tests (3 df) : <AGE50, ACPT2, PACKCAT1>

TYPE OF TEST	STATISTIC	P-VALUE
Likelihood Ratio	47.803	0.000
Wald	33.652	0.000
Exact (Conditional Scores)	46.188	0.000

Step Two:

Model: BESTACPT =ACPT2+PACKCAT1
Stratum: <Unstratified>
Weight: <none>

Number of Observations: 200
Likelihood Ratio Statistic (3 df): 61.573

	INFERENCE	<-------------PARAMETER ESTIMATION------------>			P-VALUE
TERM	TYPE	ODDS RATIO	S.E	95.0 % CONF. INTERVAL	2*1_SIDED
ACPT2	Asymptotic	32.819	NA	9.948 - 108.273	0.000
	Exact	29.432	NA	8.538 - 124.931	0.000
PACKCAT1	Asymptotic	0.720	NA	0.155 - 3.339	0.675
	Exact	0.725	NA	0.106 - 3.934	0.988
CONST	Asymptotic	0.158	NA	0.074 - 0.337	0.000

Tests (2 df) : < ACPT2, PACKCAT>

TYPE OF TEST	STATISTIC	P-VA LUE
Likelihood Ratio	47.642	0.000
Wald	33.702	0.000
Exact (Conditional Scores)	46.108	0.000

Step Three:

```
 Model: BESTACPT =ACPT2
 Stratum: <Unstratified>
 Weight: <none>

 Number of Observations: 200
 Likelihood Ratio Statistic (2 df): 59.413
------------------------------------------------------------------------------------------
               INFERENCE    <------------PARAMETER ESTIMATION----------->    P-VALUE

   TERM          TYPE       ODDS RATIO    S.E.   95.0 % CONF. INTERVAL     2*1_SIDED
------------------------------------------------------------------------------------------
   ACPT2      Asymptotic      32.510       NA      10.329 - 111.1921          0.000
                 Exact        31.959       NA       9.209 - 136.375           0.000

   CONST      Asymptotic       0.148       NA        0.073 - 0.297            0.000
------------------------------------------------------------------------------------------

Tests (1 df): < ACPT2>
    TYPE OF TEST                             STATISTIC                  P-VALUE
------------------------------------------------------------------------------------------
     Likelihood Ratio                          47.607                    0.000

     Wald                                      33.774                    0.000

     Exact (Conditional Scores)                46.017                    0.000
------------------------------------------------------------------------------------------
```

Unconditional Logistic Regression of the Secondary Data Set: Model Building Steps

Method: Forward Stepwise (LR) (Statistical Software: SPSS Version 9.0)

Step One:

Total number of cases: 200 (unweighted)
Number of cases included in the analysis: 200

Initial -2 Log Likelihood: 276.759
* Constant is included in the model

-- ------Variables in the Equation---

Variable	B	S.E.	Wald	df	Sig.	R	Exp(B)	95% CI for Exp(B)
Constant	0.1001	0.142	0.500	1	0.478	NA	NA	NA

---Variables not in the Equation---

Residual Chi Square (3 df): 49.144   Sig. = 0.000

Variable	Score	df	Sig
ACPT2	37.073	1	0.000
TECH	14.617	1	0.000
ATSREPRO	8.913	1	0.003

Step Two:

Variable Entered on Step Number 2: ACPT2

- 2 Log Likelihood: 238.432
Chi -Square (df 1): 38.327
Significance: 0.000

-- ------Variables in the Equation---

Variable	B	S.E.	Wald	df	Sig.	R	Exp(B)	95% CI for Exp(B)
ACPT2	1.846	0.314	34.520	1	0.000	0.343	6.335	3.422 - 11.727
Constant	-1.046	0.228	21.050	1	0.000	NA	NA	NA

---Model if Term Removed---

Term Removed	Log Likelihood	-2 Log LR	df	Significance of Log LR
ACPT2	-138.379	38.327	1	0.000

---Variables not in the Equation---

Residual Chi Square (2 df): 14.902   Sig. = 0.000

Variable	Score	df	Sig
TECH	11.911	1	0.001
ATSREPRO	6.447	1	0.011

```
Step Three:

 Variable Entered on Step Number 3: TECH

 - 2 Log Likelihood: 226.513
 Chi -Square (df 1): 50.246
 Significance: 0.000

---------------------------------------- ------Variables in the Equation---------------------------------------------

 Variable      B      S.E.      Wald     df    Sig.        R       Exp(B)   95% CI for Exp(B)
-----------------------------------------------------------------------------------------------------
 ACPT2       1.841   0.326    31.895     1    0.000     0.329     6.302     3.327 - 11.939

  TECH       1.102   0.326    11.458     1    0.001     0.1849    3.011      1.591 - 5.701

Constant    -2.705   0.562    23.125     1    0.000      NA        NA            NA

----------------------------------------------Model if Term Removed----------------------------------------------

Term Removed     Log Likelihood     -2 Log LR      df    Significance of Log LR

ACPT2               -130.978          35.444        1            0.000

TECH                -119.216          11.918        1            0.001

--------------------------------------------Variables not in the Equation---------------------------------------------

Residual Chi Square (2 df): 14.902    Sig. = 0.000

      Variable          Score       df      Sig

    ATSREPRO            3.231        1     0.072

No more variables can be deleted or added.
```

Exact Logistic Regression of the Secondary Data Set: Model Building Steps

Method: Backward Elimination (Statistical Software: LogXact Version 2.1)

Step One:

Model: BESTACPT =FEVPCTC+ATSREPRO+ACPT2+TECH
Stratum: <Unstratified>
Weight: <none>
Number of Observations: 200
Likelihood Ratio Statistic (5 df): 58.491

	INFERENCE	<------------PARAMETER ESTIMATION------------>			P-VALUE
TERM	TYPE	ODDS RATIO	S.E	95.0 % CONF. INTERVAL	2*1_SIDED
FEVPCTC	Asymptotic	0.153	NA	0.0245 - 0.957	0.050
	Exact	0.159	NA	0.0126 - 1.226	0.088
ATSREPRO	Asymptotic	2.105	NA	1.018 - 4.351	0.047
	Exact	2.080	NA	0.955 - 4.600	0.067
ACPT2	Asymptotic	6.433	NA	3.330 - 12.427	0.000
	Exact	6.178	NA	3.107 - 12.741	0.000
TECH	Asymptotic	2.774	NA	1.419 - 5.425	0.003
	Exact	2.711	NA	1.332 - 5.640	0.005
CONST	Asymptotic	0.0522	NA	0.0163 - 0.167	0.000

Tests (4 df) : <FEVPCTC, ATSREPRO, ACPT2, TECH>

TYPE OF TEST	STATISTIC	P-VALUE
Likelihood Ratio	57.991	0.000
Wald	41.513	0.000
Exact (Conditional Scores)	52.202	0.000

Step Two:

Model: BESTACPT =ATSREPRO+ACPT2+TECH
Stratum: <Unstratified>
Weight: <none>
Number of Observations: 200
Likelihood Ratio Statistic (4 df): 53.965

	INFERENCE	<------------PARAMETER ESTIMATION------------>			P-VALUE
TERM	TYPE	ODDS RATIO	S.E	95.0 % CONF. INTERVAL	2*1_SIDED
ATSREPRO	Asymptotic	1.918	NA	0.938 - 3.922	0.074
	Exact	1.898	NA	0.881 - 4.146	0.108
ACPT2	Asymptotic	6.208	NA	3.258 - 11.830	0.004
	Exact	6.013	NA	3.062 - 12.174	0.006
TECH	Asymptotic	2.641	NA	1.372 - 5.086	0.000
	Exact	2.598	NA	1.296 - 5.297	0.000
CONST	Asymptotic	0.052	NA	0.016 - 0.167	0.000

```
Tests (3 df) : < ATSREPRO, ACPT2, TECH>
    TYPE OF TEST                        STATISTIC                   P-VALUE
-----------------------------------------------------------------------------------
      Likelihood Ratio                    53.465                      0.000

      Wald                                40.641                      0.000

      Exact (Conditional Scores)          49.144                      0.000
-----------------------------------------------------------------------------------

Step Three:
  Model: BESTACPT =ACPT2+TECH
  Stratum: <Unstratified>
  Weight: <none>

  Number of Observations: 200
  Likelihood Ratio Statistic (3 df): 50.746
-----------------------------------------------------------------------------------
               INFERENCE   <-------------PARAMETER ESTIMATION------------>   P-VALUE

   TERM          TYPE      ODDS RATIO   S.E.   95.0 % CONF. INTERVAL   2*1_SIDED
-----------------------------------------------------------------------------------
   ACPT2      Asymptotic     6.302       NA       3.327 - 11.939         0.000
                Exact        6.171       NA       3.155 - 12.443         0.000

   TECH       Asymptotic     3.011       NA       1.591 - 5.701          0.001
                Exact        2.977       NA       1.514 - 5.975          0.001

  CONST       Asymptotic     0.067       NA       0.022 - 0.201          0.000
-----------------------------------------------------------------------------------
Tests (2 df): < ACPT2, TECH>
    TYPE OF TEST                        STATISTIC                   P-VALUE
-----------------------------------------------------------------------------------
      Likelihood Ratio                    50.246                      0.000

      Wald                                39.618                      0.000

      Exact (Conditional Scores)          46.542                      0.000
-----------------------------------------------------------------------------------
```

Appendix 18:

Comparison of Spirogram Acceptability

Between Consecutive Testing Sessions

in the Primary Data Set

Table 18.1 Comparison of Spirogram Acceptability[†] Between Consecutive Testing Sessions in the Primary Data Set

Order In Which Participants Received Testing By Each Technician	Acceptability of Session 2		
	Acceptable n	Not Acceptable n	p-value*
Session 1: Minimally Trained Technician Session 2: Highly Trained Technician			
Acceptability of Session 1			
Acceptable	8	8	0.210
Not Acceptable	15	16	
Session 1: Highly Trained Technician Session 2: Minimally Trained Technician			
Acceptability of Session 1			
Acceptable	11	7	0.134
Not Acceptable	15	20	

* p-value based on McNemar's Test

[†] Acceptable spirograms were judged accordingly by all Respiratory Epidemiologists and Pulmonaıry Specialists.