Evaluating training effectiveness in the malaysian public service

Anesee Ibrahim

Evaluating Training Effectiveness in the Malaysian Public Service

Anesee Ibrahim

A Thesis presented for the degree of
Doctor of Philosophy

Statistics and Probability
Department of Mathematical Sciences
University of Durham
England

May 2008

Dedicated to

My family: Anisah, Ahnaf, Arina, Afnan, and Aisyah.

Evaluating Training Effectiveness in the Malaysian Public Service

Anesee Ibrahim

Submitted for the degree of Doctor of Philosophy
May 2008

Abstract

The National Institute of Public Administration (INTAN) is the main training institute for the Malaysian Public Service. It plays an important role in the development of the human resources in the Malaysian public sector. However, the current method of the evaluation of the training programmes are carried out at the *reaction* level of the Kirkpatrick's model of evaluation (Kirkpatrick, 1967), giving very little indication of the effectiveness of the training programmes. The main purpose of this study thus is to develop a tool to measure *learning*, which would indicate effectiveness by examining whether there have been any changes in the level of *knowledge, skills,* or *attitude* of the training participants. Data from a total of 760 training participants are used in this study, and several different statistical analyses are carried out, namely *reliability tests, structural equation modeling (SEM), principal variables, tests of differences*, and *analysis of covariance (ANCOVA)*. Besides the main Learning Questionnaire, the Course Experience Questionnaire (CEQ) (Ramsden, 1987) and the General Health Questionnaire (GHQ) are also used. Findings indicate that the LQ needs to be modified. Model fits of the other two questionnaires are also found to be not very good. Work in this thesis continues with methods of comparing models graphically, based on the eigenstructures of the covariance matrices. The Learning Model which forms the basis of the Learning Questionnaire is applicable to other training institutes with appropriate modifications, while the statistical method of comparing eigenstructures proposed here is applicable to the general multivariate data analysis.

Declaration

The work in this thesis is based on research carried out at the Statistics and Probability Group, the Department of Mathematical Sciences, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it all my own work unless referenced to the contrary in the text.

Acknowledgements

I am most grateful to David Wooff, who served as my supervisor, without whose help this would not have been possible. To my wife, Anisah Mahmood, who always reminded me that we were into this together.

My special gratitude to Dr Malek Shah Mohd Yusuf, the then Director of INTAN, for giving his blessing for this research; Mr Yasin Salleh, the then Director of the Human Resource and Management Services Division of the Malaysian Public Service, who without hesitation gave the green light for the collection of data at the PSD main complex at Putrajaya; other people who directly and indirectly helped in the collection of data both at INTAN and Putrajaya: Ms Munirah A. Bajanuddin, Mr Razali Malek, and Mr Hadzir Md Zain.

Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The National Institute of Public Administration is the main training institute for the Malaysian Public Service. Better known by its Malay acronym INTAN which means 'diamond', the institute was established to develop human resources in the Malaysian public sector, through the designing and running of quality training programmes [70]. Latest statistics show that more than forty thousand personnel come every year to either the main campus or one of its six regional campuses to attend courses, seminars, conferences and other training related activities [40]. As the main training institute, INTAN designs and manages training programmes for the Malaysian public service personnel of all levels, who come from all federal agencies, state governments, as well as statutory bodies. INTAN is also one of over thirty centres in Malaysia that run the Malaysian Technical Cooperation Programme (MTCP), where it plays host to international participants from over 130 developing countries every year.

INTAN was established in 1972 to replace the Training Centre for Government Officers [68]. The latter was opened in 1963 to replace the first ever training centre which had been established at the coastal town of Port Dickson in 1959. INTAN's establishment in September 1972 was the result of a report by the Administrative Modernisation Unit, Prime Minister's Department, which proposed an establishment of such an institute to focus on management training for government officers.

Under the Ninth Malaysia Plan for the period of 2006 - 2010, the Malaysian government had further identified two important roles for INTAN [62]. First, INTAN would continue to provide advanced executive training for top-level government officers. For this purpose INTAN would need to develop cooperation with qualified experts and established institutes which could help in providing specific programmes. Secondly, in order to keep on improving the public service delivery system, INTAN and the Public Service Department would play an important role in the process of selecting and training the officers to ensure they were capable of contributing to the achievement of these objectives.

1.1 INTAN Trainers and Training

All INTAN officers have to go through a training-of-trainers (TOT) programme the first time they are posted to INTAN. In this programme, these officers are given basic training to be full-fledged trainers. At the end of the programme, each officer will do a personal presentation of a topic of his or her choice, during which he or she is evaluated. He or she is then recommended to be an INTAN trainer upon successful completion of this programme. Besides this required initial training, INTAN management also promotes continuous learning. Each officer is required to attend suitable training programmes, whether within or outside of INTAN, to a total of at least 14 days every year. The requirement for supporting staff is less, which is a minimum of 7 days a year.

Most courses in INTAN use classroom style teaching. The number of participants usually ranges from twenty to forty in each class, but INTAN also runs seminars and talks for over three hundred participants. In classroom teaching, the trainer gives his or her input to the class in three 2-hour sessions a day. In between the sessions participants are allowed breaks. Standard teaching equipments available in the classrooms include whiteboards, computers and projectors, and overhead slide projectors. It is also very common that participants are engaged in group discussions as appropriate in adult training methodology.

Training programmes at INTAN can be divided into two types; mandatory and optional. Mandatory courses are mostly those targeted at officers who are just being promoted to a higher grade. Confirmation to this higher grade is usually subject to successful attendance in this type of programmes. Optional courses are scheduled programmes for which participants apply directly to INTAN. INTAN acts as the provider of these programmes; potential participants are expected to choose and apply after considering their own plans of career development, usually upon agreement with their heads of departments.

I started my job at INTAN in March 1994. As an Administrative and Diplomatic Officer in the Malaysian Public Service, I could technically be posted to any one of the 26 ministries or the many more government departments. Due to the fact that I was the only candidate at that time with a first degree in Statistics, I was posted here. The institute was always short of personnels in the unit that runs quantitative training programmes.

In the quantitative unit, we run several different training programmes related to research methodologies and statistical data analyses. There are 5-day programmes, 10-day programmes, which are quite typical of most INTAN programmes, and a 3-month programme. The long programme is run only once a year, while the shorter ones are run between 2 to 4 times a year. In the quantitative unit I was known as a Project Officer, and the main part of my responsibilities was teaching data analysis and research methodology topics. After about ten years at INTAN, I decided to apply for scholarship to do a Phd. The Training Division of the Public Service Department approved funding for a scholarship for research related to the effectiveness of INTAN's training programmes, hence this thesis.

1.2 Background to the Problem

As the main player in the human resource development plan of the Malaysian Public Service, the management of INTAN has to have a very good idea of how well the institute is doing. They need to know whether the courses and other training programmes are achieving their objectives and consequently whether INTAN is having the desired impact on the personnels it trains. They also need to know whether the thousands of public servants who come every year really benefit from the courses, learn the knowledge and skills and use the knowledge and skills gained to be more efficient and effective in their jobs. INTAN needs to know the *effectiveness* of its training programmes.

Training programmes that are effective have significant impact on the participants with regards to these three aspects of learning: *attitude*, *skills* and *knowledge*. These aspects of training impact are recognised by the Government of Malaysia as the backbone of a public service with an effective and efficient delivery system [62]. *Attitude* refers to the way a person thinks and feels about something; *skills* are what a person needs to do something well; while *knowledge* refers to the information, understanding and skills gained through education or experience [1]. In the context of INTAN, its training programmes should be effective in developing the right attitude among the training participants. Their levels of skills and knowledge should also be increased by this training intervention, moulded in such a way that they are capable of delivering their duties to the expected quality and quantity.

Currently training programmes at INTAN are evaluated after they end. Participants are asked to evaluate on a scale of 1 - 7 several items that represent among others the aspects of (i) achievement of objectives, (ii) perceived effectiveness, (iii) benefits of the programme, and (iv) satisfaction of the participants on the contents, teaching techniques and overall management of the programme. Figure 1.1 shows the aspects covered under the current evaluation model. Ovals represent the factors being measured, while rectangles represent the measurements.

In the current evaluation procedure at INTAN, two arithmetic means are among the output produced: one for the overall programme and the other for each of the teaching staff. These means are taken to reflect the overall satisfaction of participants towards the programme they have just attended.

As a quality control procedure, overall scores for both the programme and the individual teaching staff are checked to see whether they meet INTAN's quality objectives standard [69]. If any of the scores is lower than 5.3, the managers of the programme are expected to investigate and provide a report to the management.

Being simply a customers' satisfaction gauge, the current evaluation does not provide a direct measure of the effectiveness of the training programme. Satisfied customers may very well be the result of other unrelated factors, such as interesting

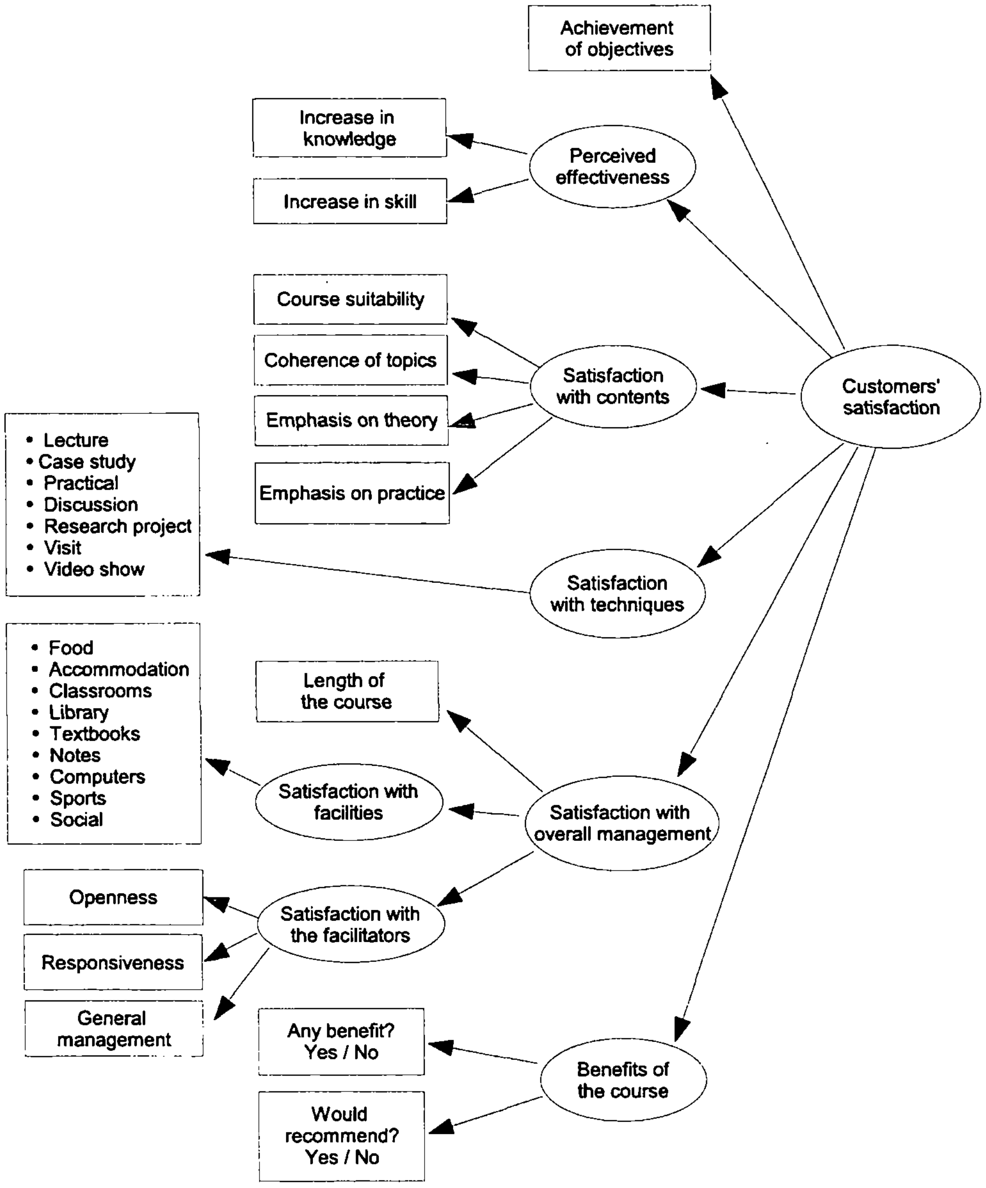


Figure 1.1: INTAN's programme evaluation model.

training programmes, good hospitality including food and lodging, and a peaceful enviroment. These would certainly result in *assessment contamination*, a term used by Rae [76]. At the same time the impact on attitude, skills or knowledge as the aspects of learning is not measured, resulting in the effectiveness of the training programme not being evaluated. Consequently, the performance of INTAN as a training institute is not clearly known.

Referring to the model of INTAN's evaluation (Figure 1.1 on page 5), the highest level factor, *customers' satisfaction*, is not explicitly analysed. All but one of the six aspects evaluated are measured on 7-point scales. The arithmetic mean calculated over the five aspects is the value taken to indicate the *performance* of the training programme.

In this model, *perceived effectiveness* is a factor which is indicated by two variables; (i) increase in knowledge and (ii) increase in skills. Having just one indicator for each aspect of learning (skills and knowledge), this scale is not expected to cover a significant domain of learning as good as a scale with multiple-item indicators. A construct (in this case; skills or knowledge) would be better measured by combining the results from a number of measures, than by taking only one individually [67].

1.3 Goals of Study

This study has several goals. First is the development of a new tool to measure training 'effectiveness'. The concept of effectiveness here is seen from the perspective of training participants as to how much learning they get from attending the training programmes. A model for *learning* is developed, and a tool (questionnaire) that attempts to measure *learning* is then designed and built. Data from the questionnaire are tested for the fit of this Learning model. The second goal of this study is to develop methods that could be used in analysing multivariate data such as data from this study. The third goal is to report the findings, particularly from the analyses of differences on the data of the three questionnaires: the General Health Questionnaire, the Course Experience Questionnaire and the Learning Questionnaire.

1.4 Statistical Methods Used

This study uses five different statistical methods on the Learning data. The report of the first method is in Chapter 5, where the reliability of the scales are examined. Here, Cronbach's coefficient alpha is used, along with alpha*, which indicates reliability value had there been only two items in the scale. In this chapter the scales are also checked for normality in their distributions.

In Chapter 6, models of the questionnaires are tested for the fit to the data. The method used is Structural Equation Modeling (SEM), which is one approach to Confirmatory Factor Analysis (CFA). In the chapter following this, Principal Variables analysis is used to examine the contribution of each of the items of the scales to the overall variance of the data. As a dimension reduction approach, this method has an advantage over Principal Components analysis, because by using the PVA, once the principal variables are identified the rest of the items could be discarded.

The tests of differences in Chapter 8 use Welch's two sample t-test and Wilcoxon test, or Analysis of Variance (ANOVA) and Kruskal-Wallis test. The first two are used to compare the means of two different levels of demographic factors, and the second two are used to compare the means of three or more levels of demographic factors.

The Analysis of Covariance (ANCOVA) is the next method employed, but only on the Learning data. The objective is to examine whether the demographic factors are associated with the posttest Learning, after adjusting for the pretest scores. The analyses are done on each of the Learning (LQ) subscales, namely the knowledge, the application, and the importance.

The last method in Chapter 10 compares the structural equation models of two hypothesised variance structures. The objective is to identify which of the two covariance structures fit the data better than the other. This method involves solving the eigenstructure problem of a compound matrix of the two covariances.

1.5 Outline of Thesis

This thesis consists of eleven chapters. The first chapter introduces the background problem and the goal of this study. In addition, it also introduces briefly the methods of analysis used in the study.

Chapter 2 is Literature Review, comprising discussions about training and evaluation of training. Here I introduce Kirkpatrick's levels of learning outcomes, the model used as the basis of this study. Other evaluation models are also presented, followed by several methods and designs of analysis specific to evaluation of training.

The third chapter is Methodology, which starts with the measurement intentions of this study, followed by the study design. Next is a discussion on the population and sample, with the details of the questionnaires following right after that. Explanations on the methods of analyses used are also in this chapter.

In Chapter 4, the discussions focus on the development of the Learning Questionnaire (LQ). The main parts of this chapter are the Learning model, which constitutes the backbone of the LQ, the development of the questionnaire itself, and the evaluation of the questionnaire. In the first part of this chapter, the specification of the Learning model under the SEM is introduced.

In Chapters 5 through 9 are the reports of the five statistical analyses. The analyses are (i) analyses of reliability (Chapter 5), (ii) Confirmatory Factor Analysis (Chapter 6), (iii) Principal Variables analysis (Chapter 7), and (iv) tests of differences and Analysis of Covariances (ANCOVA) (Chapter 8). Reports on the structural equation models comparison are presented in Chapter 10. This thesis closes with a chapter called Conclusions and Discussions. In this chapter, the results of all analyses are summarised, along with some recommendations about using the questionnaires in INTAN.

1.6 Special Names

This study makes use of three main measurement scales, eight subscales, several different statistical analyses, several different factors, and a few special names. To make reading easier, the commonly used terms are either written in their initials, underlined, or emphasized. In the Table 1.1 below are a list of scales, subscales, and special names which are used in the thesis in their initials. In Table 1.2 are the scales or factors which are presented in the thesis underlined. The words 'construct', 'factor', and 'scale' are used interchangebly in this thesis. Demographic factors are *emphasized*, whenever they are used as factors, mainly to be found in Chapter 8.

Table 1.1: Scales, subscales and special names in the thesis.

Initial	**Meaning**	**Indication**
AA	Appropriate Assessment	Subscale of CEQ
ANCOVA	Analysis of Covariance	Statistical analysis
ANOVA	Analysis of Variance	Statistical analysis
AW	Appropriate Workload	Subscale of CEQ
CEQ	Course Experience Questionnaire	Main scale
CFA	Confirmatory Factor Analysis	Statistical analysis
CG	Clear Goals	Subscale of CEQ
GHQ	General Health Questionnaire	Main scale
GS	Generic Skills	Subscale of CEQ
GT	Good Teaching	Subscale of CEQ
INTAN	The National Institute of Public Administration	Special name
LQ	Learning Questionnaire	Main scale
SEM	Structural Equation Modeling	Statistical analysis

Table 1.2: Scales and subscales/factors in the thesis.

Scale / subscale	**Indication**
Appropriate assessment	Subscale of CEQ
Application	Subscale of LQ
Appropriate workload	Subscale of CEQ
Course experience	Main scale of CEQ
Clear goals	Subscale of CEQ
General health	Main scale of GHQ
Generic skills	Subscale of CEQ
Good teaching	Subscale of CEQ
Importance	Subscale of LQ
Knowledge	Subscale of LQ
Learning	Main scale of LQ
Reaction	Factor of CEQ

Chapter 2

Literature Review

There is a lot of literature on training. Because the scope of training is very wide, I had to somehow limit our searches to only those that are either related to public service, discussed the evaluation of training, or better still, a combination of both. However we found that the availability of literature discussing the evaluation of training in the public service is very limited. In this chapter, we start by presenting models of training, followed by a section on reviewing the models. This is followed by discussions on measuring changes, methods and designs of analysis, assessments, and model and indicators.

Training, in general, is without doubt very important for developing human resources in any organization. Richard McBain [63] concluded training activities as:

> *...one of the most pervasive methods for enhancing the productivity of individuals and communicating organizational goals to new personnel.*

However, McBain then noted that consistent training evaluation was rare and many organizations did not know how their training programmes impacted performance. In general he also observed that research into training effectiveness was limited, both in terms of the types of training interventions and the evaluation methodologies [63].

2.1 Models of Training

Reid, Barrington and Kenney [78] gave some details on questions that could be asked in evaluating a particular training programme. They also quote Whitelaw [94] and Hamblin [38] in listing the following five levels of evaluation, which resembles Kirkpatrick's evaluation model (a discussion of which follows immediately):

Level 1 - reactions to content and methods.

Level 2 - learning attained.

Level 3 - job behaviour after training.

Level 4 - effect on trainee's department.

Level 5 - ultimate level, which is the well-being of the organization.

Compared to Kirkpatrick's model, Reid *et al.* included an extra level, namely Level 4. This inclusion implies that trainees benefit their departments first, before the larger organisation can benefit from their training.

In evaluating the effectiveness of training programmes, the most commonly used model is a framework of Kirkpatrick [87]. Kirkpatrick introduced four steps of evaluation: (1) reaction; (2) learning; (3) behaviour; and (4) results [52].

The first level, reaction measures the feelings of the attendees, answering the question, "How well did they like the programme?". According to Kirkpatrick, how people feel about a programme is very important; he quoted that some decisions by top management were frequently made on the basis of just one or two comments. Furthermore, training participants who enjoy the programme are more likely to obtain maximum benefit from it. The second level is learning, and just like the first level, evaluating learning should also be quantitative. Learning referred to principles, facts and techniques learned. Kirkpatrick suggests strongly the use of a before-and-after approach for this level, as well as the use of a control group.

The third step is behaviour. The question to be answered is 'what changes in job behaviour have resulted from the programme?'. Clearly behaviour is more difficult to evaluate than reaction, and there is also a big difference between knowing the principles and techniques, and actually using them on the job. Again Kirkpatrick suggests the use of a before-and-after approach, as well as a control group. Post training evaluation should be carried out at least after three months, to give the training participants the opportunity for putting it into practice.

The last step are the results to the organisation. Examples given for this level include reduced turnover of staff and decreased costs. This is extremely difficult to evaluate due to the difficulty in determining the real improvement due to training. Kirkpatrick recommends that most evaluations be made on just the first three steps.

Based on Kirkpatrick's model, the current practice of evaluating INTAN's training programmes is equivalent to the first level, which is measuring the *reaction* of the participants towards the programme they have just attended. The three higher levels of the evaluation model, namely the *learning, behavior* and *results*, which would better answer the questions of training effectiveness are not measured.

Bramley [13] explained two major models of training, the individual training model (Figure 2.1) and the increased effectiveness model. The first model which focuses on individuals has greatly influenced training for trades and technical training. The process involved in training based on this model is encouraging the trainees to learn something useful and expect them to find uses for the learning.

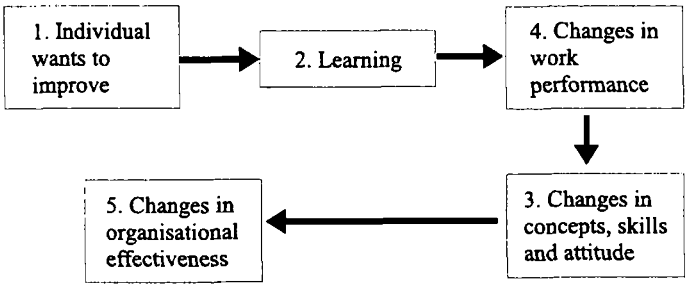

1. Individual wants to improve
2. Learning
4. Changes in work performance
3. Changes in concepts, skills and attitude
5. Changes in organisational effectiveness

Figure 2.1: Individual training model.

The second model is based on changing effectiveness, rather than on educating individuals. In this model, aspects of job situation other than skills of the people are considered when defining resources. Learning impact is measured by changes in job performance, not by changes during training (as in the previous model). According to Bramley [13], it is more appropriate for the kind of work where people have some descretion about what they do, or the ability to negotiate priorities.

Mahapatra and Lai [61] proposed a framework for evaluating end-user training in information technology (IT). It has two dimensions, the evaluation dimension and the evaluator dimensions. The former dimension has five levels: (1) technology, (2) reaction, (3) skill acquisition, (4) skill transfer, and (5) organizational effect. In this dimension, levels 2 to 5 are fully compatible with the four levels of Kirkpatrick's model [52]. The second dimension in this proposed evaluation framework is the evaluator dimension. It consists of the training providers, the trainees, and the managers. Each group of people in this dimension must evaluate factors at various levels of the first dimension.

Bramley [13] also emphasized that the impact of training must not stop at the trainees' attitude or behaviour only. Changes in these two aspects should also result in greater effectiveness in the organization. At the trainee's individual level, the impact of training is measured in three aspects - knowledge, skills, and attitude. They are multifaceted, and many of them integrated. Bramley also believes that learning affects the whole person and any increase in knowledge or skills would usually result in different attitudes to some aspects of the work. According to Bramley, evaluation is an integral part to the training cycle. It has a key role of quality control by providing feedback on the effectiveness of methods, achievement of objectives, and whether the original needs of the organization and the trainees have been met.

To gauge the success of a training programme, Leslie Rae [76] divided the discussion into two aspects; the training itself and the effect of the training on the work. There are many aspects of training that could be evaluated or validated. Rae listed a few possible aspects as follow:

Content of training The main question asked here is whether the contents are relevant with training needs, or are they really up to date.

Method of training Different subjects sometimes require different methods of training to be effective. So do the participants, who might have different learning styles.

Amount of learning How much do the participants learn and how much of it is actually useful. Parts of the material could be revision and not totally new.

Trainer skills The trainers need to have the right skills to do the job effectively. Just as important is the right attitude of the trainers.

Length and pace of the training This aspect looks at whether the length of the programmes is appropriate, and whether the programmes are going at an acceptable pace to all participants.

Objectives Satisfying the declared objectives of the organisation is the most important part of training. In addition to that, managers should also look into the possiblility of satisfying participants' personal objectives.

Omissions There is always this possibility that some essential aspects are ommitted from the content. Conversely, some materials which are not essential could have been included.

Learning transfer Training is not of much use if the learning stops in the classroom and not put into practice. It is always desirable to be able to identify factors that deter or assist the transfer of learning into action.

Accommodation Though not directly related to training, accommodation and meals are nevertheless important factors to be considered.

Relevance Relevance of the total training approach is asked as a final question in a validation assessment.

Application of learning Questions that could be asked include the aspects of work directly affected by the learning event. Are there any aspects replaced or introduced? Are there any aspects that have not been applied, and if so, why?

Efficiency Have the participants become more efficient and/or more effective in work as a result of the training?

Hindsight This part asks about any amendments the participant wishes to make to his immediate outcome validation answer.

Arthur, Bennett, Edens and Bell [46] did a meta-analysis of the relationship between design and evaluation features and effectiveness of training in organizations. Their overall conclusion was that the effectiveness of training programmes is related to the training method used, the skill or task characteristic trained, and the choice of training evaluation criteria. However time intervals between end of training programmes and evaluation done is not related to the observed effectiveness. Neither is needs assessment, though this was cautiously concluded from just 4 data points.

2.2 Review of Models of Training Evaluation

The Kirkpatrick model is the best-known and most widely used framework for classifying evaluation [87]. It is simple, pragmatic and easily comprehended, and thus makes sense to organisations to adopt as a model or framework for training evaluation. In recent years, there have been criticisms too. Tamkin *et al.* [87] reviewed some of the criticisms, like those of Bernthal [9], Alliger and Janak [3], and Holton [39]. Tamkin *et al.* however pointed out that one of the most common criticisms was based on a misunderstanding that the levels had been arranged in ascending value of information, giving the impression that *result* data were more important than *reaction*.

Tamkin *et al.* also listed some of the models which they called *Kirkpatrick plus*. These are the models proposed over the last 40 years and had used Kirkpatrick's framework as the basis. These are the Five Level Approach (Hamblin, [38]), the Organisational Elements Model (Kaufman, Keller and Watkins, [48]), the Indiana University Approach (Molenda, Pershing and Reigeluth, [65]), the Five-level ROI Framework (Phillips, [71], [72]), the KPMT model (Kearns and Miller, [49]), and the Context, Input, Reaction, Outcome Approach (Warr, Bird and Rackam, [91]). Less well known models but were deemed worth mentioning included those by Brinkerhoff [14], Bushnell [16], Sleezer, Cipicchio and Pitonvak [83], and Fitz-enz [28]. Overall, these models suggest expansion, both before assessing reactions and after evaluating results of the training.

On models alternative to Kirkpatrick's, Tamkin *et al.* divided them into two categories: those that focused on the purpose of evaluation, and those that provided alternative measures. Summary of the models are listed below:

2.2.1 Models that Focus on the Purpose of Evaluation

Responsive evaluation: Pulley [74] The objective of the evaluation should be to provide evidence so that key decision-makers can determine what they want to know about the programme.

Educational evaluation: Stufflebeam, D., W. Foley, and others [86] Developed for used in an educational context, distinguishes four types of evalution - context, input, process and product.

Newby [66] Evaluation can be done in many different contexts: within the training event, in the workplace, in the context of performance measures, and also using criteria not related to the workplace, such as societal, moral, political or philosophical.

Evaluative enquiry: Preskill and Torres [73] Evaluation is a learning process; connected to the organisation's mission and strategic plans.

2.2.2 Models Using Different Measures

The learning outcomes approach: Kraiger, Ford and Salas [55] Suggested the need to distinguish the three different types of outcomes - Cognitive, Skill-based and Affective. This can be done by viewing the instructional objectives through different ‘lenses‘.

The Balanced Scorecard: Kaplan and Norton [47] Aims to balance business management by measuring across four different perspectives - finance, customers, internal business processes, and learning and growth.

Concept Mapping and Pattern Matching: Anderson Consulting Moad [64] and Abernathy [2] developed concept mapping and pattern matching, based on the premise that managers know the skills and behaviours needed by their employees.

2.3 Measuring Changes

Bramley [13] discusses the evaluation of training effectiveness by looking at the changes in several different aspects, namely knowledge, skills, attitude and behaviour, and effectiveness of the organization.

Changes in knowledge.

Knowledge is required for anybody to do a job. According to Bramley, knowledge can be divided into 3 levels, the basis of which is that of isolated pieces of information. Examples given include the ability to recall simple lists, know simple facts, and state simple rules. A higher level is to be able to arrange pieces of information into procedures, such as how to do things, and how to order a set of actions. Higher still is the analytical ability. Essentially this is the ability to make some decisions regarding procedures or methods, after analysing them for their key elements. It is not possible to achieve higher levels without the lower levels. The functions of training therefore can be seen as (i) analyzing what is required at each level, (ii) discovering what trainees know at each level before attending the training, (iii) trying to close the gap, and (iv) evaluating the extent of them being below satisfactory job performance level at the end of training.

Changes in skills.

Bramley defined skill as the ability to perform a task well. There are four levels of skills suggested:

1. Basic ability to communicate. Examples for this level include labeling items and identifying parts.

2. Ability to do simple procedures, often with the use of instructions/notes.

3. Physically skilled actions. This involves hand-eye coordination, and requires considerable practice.

4. Judgment. Ability to evaluate whether a work done is of acceptable quality.

Bramley's definition of skills is not restricted to what it normally means, which relates only to physical skills. Ability to communicate effectively and ability to judge are crucial for managers in any organization, thus changes in skills are as important as changes in knowledge and attitude.

Changes in attitude and behaviour.

Attitude is defined to be the tendency or predisposition to behave in certain ways in particular situations. Bramley believes that it is possible to follow up changes in attitude back to the workplace, but he is doubtful if it will produce useful information. The assumption that 'changes in attitude imply changes in behaviour at work', will still be there. He suggested instead the use of a behaviour scale, which he believes is more likely to be helpful. This scale measures changes in the ways things are done.

The best demonstration of training's value is when learning translates into lasting behavioral changes. James Kirkpatrick [53][1] feels that learning transfer has not been paid enough attention compared to the other three of Kirkpatrick's levels of evaluation. Just like in a business environment, corporate universities and training departments too have focused their calculations on the final results. But he pointed out that in order to gain maximum benefit and meaning from the measures of training values, it must be done effectively.

Changes in effectiveness.

Some writers believed that the ultimate objective of training is to increase the effectiveness of the organization. However many argued that training could not be evaluated against organizational effectiveness, because changes due to training are indistinguishable from other factors, or the effort of an individual has little effect on organization as a whole. But Bramley believes that it is possible to do so, by focusing on a small part of the organization and to link improvement in performance with training inteventions.

[1] James Kirkpatrick is the son of Donald Kirkpatrick. He is a consultant and workshop conductor, and a practitioner in the field of training.

Evaluation on the impact of training on organizational productivity is rare. McBain [63] attributed this partly to the difficulties of gaining data, of separating the effects of training from those of other interventions, and the lack of suitable evaluation methods. Nonetheless he agrees that evaluation at this level is the most critical, because it can identify whether the training has met its needs.

Folley [30] emphasized the fact that measuring the effectiveness of training was not easy. However, he believes that half of the battle is won if the objectives of the training programme have been developed well. According to him, statement of objectives provides the means for evaluation of training that is valid by definition. What remains is to construct specific test items and devise the scoring system. At the individual trainee level, the question to be answered is 'How well has the trainee achieved the training objectives established earlier?'. From a collection of results from all trainees, the question becomes, 'How well has the training achieved its objectives?'. Validity of evaluation, according to Folley, referred to 'the ability to perform behaviours that appear in the objectives'. Related to that, he mentioned two main problems, (i) performance that could not be directly measured, and (ii) how to score performance. For the former, an indirect measure was used, such as a pencil-and-paper test of knowledge and perception. However, it was pointed out that sometimes there is little relationship between what a person is able to tell and what he is able to do.

According to Youmans [99], employers generally has two expectations in testing. One is what a person could do, and the other is what he will do. What a person can do is related to his ability or capability, and can be divided into three classes as follows:

1. General mental or learning ability, which includes alertness, intelligence and adaptability.

2. Achievement - proficiency in performing skills and in using general and technical knowledge.

3. Aptitude - indicated capability or potentiality for learning skills and knowledge.

What he will do refers to his emotional, temperamental and motivational attitude. Both expectations seem more appropriate for discriminative testing in the selection process, but can also be applied to evaluation.

When investigating the interrelationships among sales training evaluation methods, Leach and Liu [58] found that Kirkpatrick's four levels of evaluation were hierarchical, where higher levels could be predicted by the lower levels. Trainees with positive reactions to a programme are more likely to learn the material (level 1 → level 2). Then, trainees who acquire more knowledge are more likely to transfer learned material to the workplace (level 2 → level 3). They also divided level 4 of Kirkpatrick's model into three organizational objectives: (a) improve organizational commitment, (b) improve selling effectiveness, and (c) improve customer relations. In their study, level 3 (learning transfer) was found to be related to all three objectives. As that is the only type of evaluation that could explain level 4 (results), they agree that learning transfer in training design is of critical importance.

There are many reasons why the effectiveness of a particular training programme should be evaluated. For the Birmingham University Interprofessional Training Programme, Carpenter, Barnes and Dickinson [20] listed three main reasons:

1. there is widespread uncertainty about outcomes;
2. evidence for the effect of training is usually not strong;
3. evaluations had tended to be flawed for they had mostly been:
   (a) not independent;
   (b) short term;
   (c) had not follow through to end users;
   (d) lacking in strong methodological design.

For the evaluation, an expanded version of Kirkpatrick's four levels of evaluation based on the work of Barr, Hamick, Koppel and Reeves [8] was suggested, as follows:

Level 1 Reaction - participants' view on their learning experience and satisfaction.

Level 2a Modification of attitude/perception - between participants and towards users.

Level 2b Acquisition of knowledge/skills - refers to concepts, procedures, principles and skills.

Level 3 Change in behaviour - transferred from learning environment to workplace.

Level 4a Change in organizational practice - wider changes in the organization.

Level 4b Benefits to service users - improvements in the health and well being of service users as direct result of the educational programme.

Carpenter [19] suggested this expanded model in his discussion paper for evaluating outcomes in social work education. A work by Barnes, Carpenter and Bailey [6], in which service users were asked about desirable outcomes of professional education, was also cited. In the study it was found that service users had stressed outcomes of Level 2, instead of Level 4. Level 2 refers to attitudes, knowledge and skills, while Level 4 is the one that referred to the benefits to the service users.

In the paper Carpenter mentioned the scarcity of evaluative research on outcomes of methods in social work education. Narrative research was plentiful, but carefully designed research was rare to find. Controlled evaluation was even harder to find. This led to limited information in published accounts which became a major problem in establishing an evidence base for social work education. More and better quality evaluations were needed, but before that could be achieved, researchers must be clear on what to evaluate and how to evaluate them. As a framework, Carpenter used the work of Kraiger, Ford and Salas [55] as the main reference. He further stressed that his discussion paper emphasized the outcomes of learning and how they might be evaluated. He was, therefore, not concerned with philosophies, curriculum design, modes of learning or course content. Nevertheless, understanding of the process of learning was important.

On measuring the outomes, Carpenter [19] applied Kraiger and colleagues' [55] model to social work education. The model is an elaboration of Kirkpatrick's Level 2, distinguishing cognitive, skill-based and affective outcomes. Suggestions as to how the outcomes could be measured were also listed. A reproduction of the table in the discussion paper is presented in Table 2.1.

Table 2.1: Suggestions to measure learning outcomes - Carpenter [19]

Aspect	Dimension	Measurement
Cognitive	Declarative (verbal knowledge) Procedural (knowledge organisation) Strategic (planning, task judgement)	Multiple-choice questionnaires Concept mapping; case study Probed protocol analysis
Skills	Initial skill Compilation of skills Advanced skills (automaticity)	Self/observer ratings Observer ratings Observation
Affective	Attitudes to users; values Motivational outcomes, self-efficacy	Attitude scales Self-ratings; confidence ratings
Behaviour	Implementation of learning (and barriers)	Self-report; practice report rating scales
Impact	Outcomes for users and carers	User-defined scales; self-esteem and empowerment; measures of social functioning; mental health, quality of life, etc.

The proposal by Carpenter was specialised to social work education. He also listed four questions to be answered from the evaluations, namely:

1. Does it work? Do students learn the outcomes which educators hope they do?
2. Are they able to put their learning into practice?
3. If so, does it make a difference to the lives of service users and carers?
4. Is any particular method more effective than any other method in practice.

Robinson and Robinson [80] explained two reasons why evaluation is not entirely over after a training programmes ends. First, in order to know whether the level of knowledge or frequency of skills or behaviour had changed, it is best to have before-and-after evaluation. Secondly, objectives of the programme have to be identified in terms of training outcomes. These have to be included in the design of the programme to have the best impact.

For the evaluation, Robinson *et al.* also suggested a modification of Kirkpatrick's four levels of evaluation. It has five levels, where level 3 is divided into two types, type A and type B. Details are as follow:

Level 1 Reaction evaluation - critique; customer satisfaction index.

Level 2 Learning evaluation - quality assurance index. Questions to be answered: (i) have they learned the stated objectives? (ii) how can we be sure learning objectives are accomplished regardless of who the instructor is.

Level 3 (Type A) Behaviour or skill application evaluation - are they using on the job what they have been taught?

Level 3 (Type B) Evaluation of nonobservable results - includes mental use of problem-solving technique, and commitment.

Level 4 Bottom line impact. In business, it is money. In some other sectors, customer satisfaction is the bottom line.

2.4 Method and Design of Analysis

For testing gain of knowledge, Bramley [13] suggested the following methods:

1. Open-ended questions (essay).
2. Short answer items (describe; define; determine; etc.)
3. Objective test items.
4. Multiple choice questions.
5. True/False questions.

For testing levels of skills he suggested practical tests, with two possibilities, (i) the trainee is set a task, and work inspected at the end, and (ii) the trainee is observed throughout the test, so methods that he uses can also be assessed. Suggested method for checking a change in attitude includes using semantic differential scales. Here participants are asked about their opinion on particular concepts, and they respond on seven-point scales. Frequencies or averages of the group will show their overall attitude on the concept. To identify change, this is done before and after training.

The above suggestions however are not suitable for use in INTAN. Methods to be employed in INTAN must not be time-consuming to complete, and must not take long for the results to come out. Methods which are very specific to certain training programmes are also not suitable. Training programmes in INTAN vary a lot in contents, objectives, focus, etc., and a method to be used must be general enough to be usable in all programmes. Practical tests are also not suitable, as most training programmes in INTAN are not technically oriented.

The other suggested method was "repertory grid"[2]. Bramley mentioned that this method was rigorous and sophisticated, with many variations to suit particular situations. In his discussion paper [19], Carpenter explains several potential research designs for evaluating outcomes in social work education. They are as follows:

Post-test only This design is the easiest and the most commonly used. However, since there is no pretest score to compare it to, the score cannot be attributed to the learning intervention. Therefore it is inadequate for evaluation purpose.

Pre-test and post-test The same questionnaire is used for two measurements and the differences are observed. The differences could be attributed to the learning intervention, but nevertheless, they could also be attributed to other factors such as "maturational effect"[3], or "Hawthorne effect"[4].

Post-test, two groups Appropriate in comparing two different groups or methods. Differences among the students are evened-out by randomization. Since measurement is only done once, there is no opportunity for contamination by practice or maturation. This enables one to test one method against another, but not to tell how much the students have learned.

[2] The Repertory Grid is an interviewing technique which uses factor analysis to determine an idiographic measure of personality. It was devised by George Kelly in around 1955 and is based on his Personal Constructs theory of personality (Source: http://en.wikipedia.org/wiki/Repertory_grid).

[3] When observed outcomes are a result of natural changes of the programme participants and not entirely as a result of the training intervention, maturation effect has taken place. This effect is generally considered as a threat to internal validity of an evaluation.

[4] The Hawthorne effect refers to a phenomenon which is thought to occur when people observed during a research study temporarily change their behaviour or performance. The term gets its name from the Hawthorne Works, where a series of experiments on factory workers were carried out between 1924 and 1932

Pre-test and post-test, two groups Here quasi-experiment[5] was adequate, without the need for random allocation. This is potentially very useful for the reasons (i) students do not have a feeling of getting a worse intervention than colleagues, and (ii) the greater sample size increases statistical power. Differences among students at time 1 could be adjusted statistically using analysis of covariance (ANCOVA). Another important factor is the trainer's competence, which might have an impact on learning outcomes.

It is not easy to get a "non-intervention" control group. Carpenter [19] suggested a waiting list control with repeated measures (Figure 2.2). Measurements were taken at three points on both groups. Group 1 received training intervention after time 1 (Tl) and other studies after time 2 (T2). Group 2 meanwhile received other studies after time 1 and received training intervention after time 2. The third measurement (T3) is made at the end of evaluation. Differences between time 2 and time 1, and between time 3 and time 2 were compared between the two groups.

This is essentially a crossover trial where all subjects receive the 'treatments'. While it has its advantages, it is also quite hard to analyse well. There will always be a washout period of the treatment received, and the subjects start on the following treatment while still being affected by the previous one.

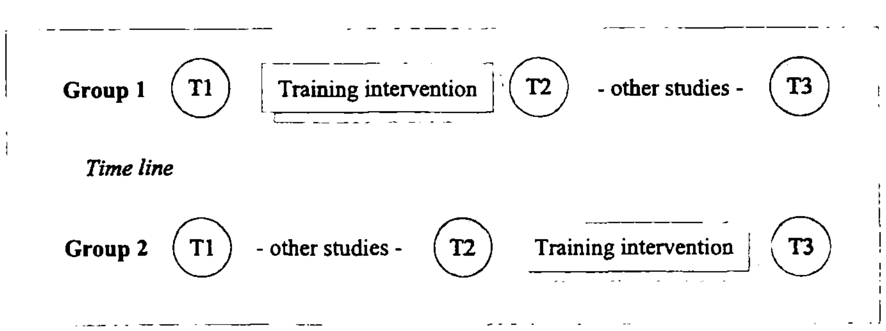


Figure 2.2: Waiting list control with repeated measures.

Time series Conclusions are based on trends before, during and after intervention.

[5] Quasi means 'almost'. Quasi-experiment refers to an experimental design where the researcher has less control over the independent variables. Inability to randomly assign participants is a usual example, as the case mentioned here.

> In a more sophisticated design, the intervention is withdrawn and subsequently reintroduced and the effects are noted.

The use of control groups is also suggested by Leslie Rae [76] in his book, *How To Measure Training Effectiveness.* Control groups become an absolute requirement when the learning to be evaluated is over a period of time. However, although the absence of a control group can affect the objectivity of the evaluation, its presence does not guarantee complete objectivity either. All of these groups ideally should be a complete match in terms of job, age, experiences, skill level, education, intelligence and many other relevant characteristics. But in reality, Rae pointed out that it was very difficult to put into practice. Anyhow, the results must be treated with some care.

If the non-training groups show similar change to that of the training group, serious doubts must be expressed about whether the training is necessary. Otherwise if substantial change is shown within the training group but not in the non-training group, the view that the change is due to the training is supported. If the result lay some way between these extremes, the necessary question to ask is whether all the training is necessary.

2.5 Assessments

Assessments at the start of the training event are necessary in order to assess any change following the training. The results of this can be used in the pre-test and post-test evaluation design. For assessing knowledge or skills, Rae [76] suggests the use of a questionnaire. The format for this questionnaire can include open answers, binary choice, true/false choice, multiple choice or short answer. All of them have their limitations, but they are useful especially when there is no pre-training information available.

Assessment of Skills.

Rae [76] warns that the assessment of skills is often very difficult, both for the tester and the tested. Many skills, particularly at management level, are subjective and cannot be assessed objectively. Validation attempts using subjective assessment are said to be of little value, but Rae continues in reminding that in spite of difficulties, *some* assessment is better than none at all. In view of this, Rae suggests the use of a self-completed questionnaire. Here individuals are asked to rate on a scale how effective they think they are in a number of aspects which are included in the training event.

Types of scales suggested included the semantic differential, a kind of scale where participants are asked to rate the aspects on a scale between opposites. The Thurstone scale[6] could be used to avoid the problems of allocating a numerical weighting to a subjective view by requiring just agreement or disagreement. A variation of this approach is to have the binary options as 'agree or agree more than disagree', or 'disagree or disagree more than agree'. The advantage is to do away with the need to have a choice between the extremes of agreement or disagreement only. Other scales suggested were the Likert scale and a ranking scale. A Likert scale is a type of psychometric response where respondents specify their level of agreement to a statement. With a ranking scale, a respondent is presented with several items simultaneously and asked to rank them.

Assessment of Attitudes.

Rae [76] acknowledges that assessment of attitudes is far more subjective than assessment of skills. There are two options: the first is self assessment by the participants, and the second is parallel assessment from the manager and the subordinates. The latter is usually biased and very weak, thus unreliable, so views of the participants themselves might be the best in this situation. The tester just had to assume that the questionnaire was completed honestly and with the maximum awareness by the participants of their own feelings. For this Rae suggests the use of a scale of ten to encourage them to assess the level as accurately as possible, or at least make them think hard about the rating. A further advantage is that it permits the result to be

[6] Thurstone scale - a way of measuring people's attitudes along a single dimension by asking them to indicate that they agree or disagree with each of a large set of statements that are about that attitude.

easily expressed as a percentage.

In many cases, assessment may be carried out during the training event. Rae [76] argues that the trainer often needs to know to what extent the training is having an effect. Immediate modification may be necessary to the approach or material if the expected changes are not taking place. The need to modify and the ability to do so, according to Rae, reflects the flexibility of both the trainer and the training event and is itself a measure of validation. For this purpose most of the tests of knowledge and skill can be used, either as a repetition or an updated test. They could also be administered formally or informally. Observation, according to Rae, plays an important part in the assessment of learning during a training event, particularly in training other than for specific skills or knowledge. Rae [76], however, warns of the danger of over testing. Completing questionnaires can become boring and tedious, and can become counter productive. If this is to become a concern, he suggests doing away with it and sticking with just the pre and post testings.

Validation of Assessment.

In view of the danger of over-testing, post testing is the second important validation event [76]. This is where internal validation can be differentiated from external validation. Rae considers internal validation as the assessment of the validity of the training programme itself. External validation, meanwhile, is the extent to which the learners learn from the training experience. For validating a training course, Rae further mentions the following methods: group review, end-of-course questionnaires, blank sheet review, open question validation review, feelings review, action planning, and an interview approach. The most common method is group review, even though it is most unlikely to be completely forthright and comprehensive. The other is end-of-course questionnaire, where information such as how much participants consider they have learned, how much they enjoyed the programme, and what they think of it can be gathered. More valuable assessment can be done by interviewing, but it is often too time consuming to be practical.

In validating a training programme, Rae [76] agrees that if a change (in knowledge, skills or attitude) has occurred during the training course and it is in step with the objectives, then the training is validated. As to external and internal validation, the former is thought to be more important. An effective training programme is only useful if the learners can put the learning into practice. Confirmation of attainment of skills is much more difficult than confirmation of acquisition of new knowledge. It will be even more difficult if the training programme is concerned with attitude and behavioural skills.

Contamination of assessment.

Rae also mentions about participants contaminating their individual assessment [76]. This occurs when a learner's awareness increases and his perception heightens as a result of training. Rae has a way of attempting to test the results of this possible contamination by introducing a third completion of the same questionnaire, some time after training ends [76]. According to him, this 3-test approach produces much more realistic results than the pre/post test approach. But he also admits that it does not mean that it is better. The other type of assessment contamination is the one usually referred to as 'happiness sheets' evaluation. Evaluation done immediately after a training programme has ended suffers directly from the state of the participants themselves. If it has been a very enjoyable experience, views expressed on the training can be clouded. Rae suspects that on a scale of 1 to 10, the ratings given may be 3 or 4 ratings more than what should have been realistically given [76].

Delayed assessment.

Assessment done beyond the immediate end of the course serves the purpose of allowing the emotions to subside to a more rational level. Delayed assessment also is more appropriate for evaluating the impact of learning on the work environment, as opposed to validating the training programme itself. Rae suggests a period between three to six months after the training event to do the delayed evaluation. During this period, the training programme is still reasonably fresh in the minds of the participants, and at the same time it is sufficiently distant to have allowed them to start practising the skills acquired.

Involvement of others.

Other people close to the individual being assessed can be involved in the evaluation process. This include his superiors, his peers and his subordinates, who can provide information on his work. However, Rae warns that sometimes these groups of people may not be capable of assessing the ability of the person being evaluated. Anybody used in the assessment must have a reasonable amount of contact with the person so that he has sufficient evidence on which to base an assessment. Rae listed three requirements for anybody to be eligible to provide assessment: he sees enough of the person's work; he is able to make objective judgement without bias; and he is sufficiently skilled to assess the person's skill level.

2.6 Model and Indicators

An indicator is an observed variable directly related to a latent variable, in such a way that any change in the value of the latent variable is mirrored in the value of the indicator. In dealing with latent variables which have no direct measure, there are three possible strategies, suggested in [11]:

1. **Ignore the latent variable**. This strategy has the danger that the omitted variable is likely to bias the estimates of the impact of the other included explanatory variables.

2. **Include indicators, but ignore their measurement errors**. This strategy is better than ignoring the latent variable altogether.

3. **Include indicators, and take account of errors**. This is the best choice. To take account of the measurement error, a measurement model has to be formulated.

A measurement model shows how a latent variable is measured, by relating it to its indicators. There are two types of indicators; causal and effect indicators. Causal indicators are those that influence the latent variable. Effect indicators, on the other hand, are the ones which get influenced by the changes in the latent variable (Figure 2.3).

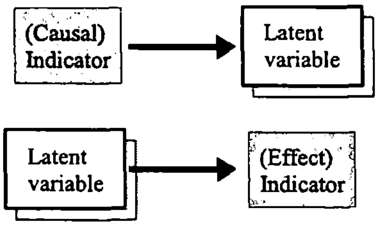


Figure 2.3: Two types of indicator variables.

In factor analysis, effect indicators are implicitly assumed, just like in much of measurement theory in social sciences [11]. However, treating all indicators as effect indicators is not correct, as they have different properties. Bollen [11] points out that estimates from a model can be biased because of incorrect classification of indicators. He also emphasises the importance of clarifying whether the indicators and the latent variables are continuous or not. Being continuous means it is best thought of as an approximation of a variable which has infinite degradation of magnitude. This is expecially true with indicators because measuring instruments have limitations that will not permit infinite degradations. It is also very much possible that a continuous latent variable is measured by a dichotomous indicator, in which case the varying degrees in the latent variable are not reflected in the choice of response in the indicator.

An increasingly popular but less common formulation is to have a continuous latent variable explained by one or more noncontinuous indicators [11]. The approach is provided by Item Response Theory (IRT), and the formulation is common in structural equation models. If what to be measured is a concept, there must be a clear theoretical definition which identifies the distinct aspects or dimensions of the concept. If a dimension is a latent variable, it is measured by indicators (Figure 2.4).

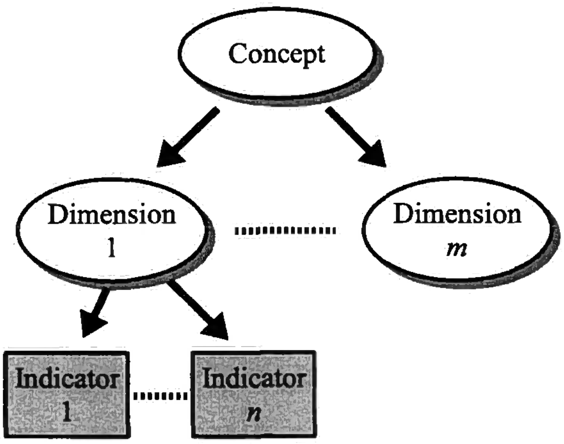


Figure 2.4: A concept is defined by m dimensions. Dimension 1 is measured by n indicators (Bollen [12]).

Chapter 3

Methodology

We will start this chapter with sections on the model and measurement, the study design, and a discussion on the concept of population and sample. After that we introduce the measurement tools, ie. the questionnaires used in this survey. An explanation of the administrative aspect of this survey comes after that, followed by a discussion on data preparation and screening. The statistical analyses carried out in this thesis are discussed in the last three sections of this chapter.

3.1 Training Evaluation Model and Measurement

Training effectiveness has been most widely explained using Kirkpatrick's four-level model. The first level, *reaction*, relates to enjoyment of training, perceived usefulness and difficulty, and general liking towards the programme. Learning is about getting new knowledge or new skills. *Behaviour* change is associated with changes in on-the-job behaviour. *Results*, which is the highest level, is measured by examining the impact of training intervention to the firm or organization.

In this study, the focus is only on the first two levels: *reaction* and *learning*. Related to the second level, Learning is the main construct of this study, and it is measured among all participants in the study group, as well as among all members of the control group. *Reaction*, on the other hand, is only measured among the study group.

As this study involves a control group which were not attending any training during the data collection period, a measure of Learning specific to a particular course was not appropriate to be used. Instead, the measurement used has to be general enough to be relevant to the control group as well, while being sensitive enough to measure the changes brought about by the training programmes attended by respondents from the study group. Furthermore, training programmes at INTAN vary widely in terms of the subject areas, the length of training, the levels of the participants, the approaches used in the programmes, and other aspects. A measure specific to a particular training programme would definitely be unsuitable to be used with training programmes as varied as those at INTAN.

Literature reviews failed to find a suitable questionnaire that could be used for measuring Learning in the context of this study. A questionnaire had to be developed specifically for this purpose. For measuring *reaction* however, the Course Experience Questionnaire (CEQ) [77] which had gone through many years of development, is used.

3.2 Study Design

INTAN needs to know whether the training programmes are effective, in which case it means that the training programmes have the desired impacts on the participants. One way to examine this is by carrying out the measurements twice, one before the training, and the second one after the training. Whatever differences there are between the two measurements may only be attributed to the training intervention if the same differences are not observed in samples not attending training. For this reason, the pretest-posttest comparison with control group is used in this study (Please refer to Figure 3.1.). This design will compare the Learning score of the course participants measured before the training starts, with a score measured after it ends. At the same time, Learning was also 'measured' among the control respondents who were not attending any training during the same time period. Changes in the scores of the two groups are then compared.

With this design, two main questions can be answered: (i) *did learning take place during the period of training?* and (ii) *could the 'learning' be attributed to the*

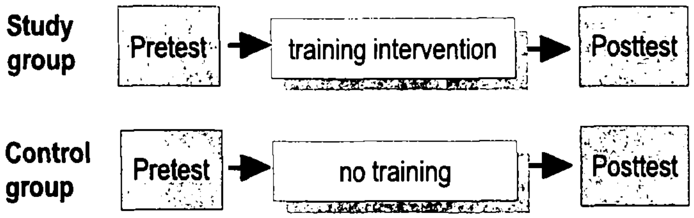


Figure 3.1: Research design used for this study.

training intervention (the programmes)? For the testing to be valid, both the study group and the control group need to be as similar as possible in terms of possible relevant variables such as sex, age, experience, etc. [76]. The only difference between the two will be the training intervention that the study group has during the study period. For this study, randomized allocation of participants to the two groups was not possible, thus a controlled trial was not an option.

Another design which is also appropriate to for this study is the "pretest and posttest two group design". This design which is also suggested by Carpenter (2005) [19] is similar to the waiting list control with repeated measures design (Figure 2.2 on page 26). While the selection of participants should be straighforward for group 1 in this research design, the same thing can not be said for group 2. The period of time between when the participant decide to attend (T1) and the start of the training programme (T2) for group 2 is sometimes very short, even 1 or 2 days are not uncommon. In cases like this, there is simply not enough time to administer the questionnaires. Besides, course participants come from all over the country, making contacting and persuading them to participate in the study quite a task. For these reasons, this design was deemed inappropriate.

3.3 Population and Sampling

The **target population** is defined as the total finite population about which information is required [7]. In this study, the target population is all course participants who come to INTAN for training. This target population is chosen because the main focus of the research is to see whether INTAN's training programmes are effective. The question to be answered is, *'Do participants that come to INTAN really benefit* *from the courses?'*. At the same time, this research will examine whether there are differences in the learning outcomes between or among the different levels of seven demographic factors.

For practical purposes, the population needs to be limited. For this study only participants from the *Management and Professional* (Prof) and the *Supporting staff group 1* (Supp) were considered. The other groups which were not considered are the *Supporting staff group 2* which is lower in position in the managerial hierarchy, and the *Premier Grade Officers*, which is at the top-most position. In the public service managerial levels, the Management and Professional group is the middle managers, positioned in between the top decision makers and the lower executives. This makes them more accessible, both while attending training at INTAN and while at work. Hence, the **study population** for this study is the Management and Professional group and the Supporting staff group 1 attending courses at INTAN. The same two groups who were not attending any courses during the study period were the study population for the control group. The other limitation is that only participants attending training at the INTAN main campus were considered. The long travelling time and the high travelling cost needed for studying participants at INTAN branch campuses made it not very practical.

The participants of this study were selected with the help of their course coordinators. A coordinator is the actual manager of a course. He or she is always in a direct contact with the participants, and this helped in the distribution of the questionnaires. While the course participants are the **sampling units**, the courses or training programmes are the **primary sampling units**. Thus the **sampling frame**, which is a set of sampling units [7], is made up of a list of courses designed for the Management and Professional and Supporting staff group 1, scheduled to run at the INTAN main campus within the study period.

The other set of sampling units was used to select individuals for the control group. It consisted of the Management and Professional and the Supporting staff group 1 officers who were not attending any courses during the period. The purpose of having a control group is to measure general fluctuation and variation in responses, which then will be compared to those of the study (treatment) group. Again, the target population is a long list. These officers were working in more that 26 ministries and government department and agencies in all over the country. Thus, the control group needed to be limited to a manageable limit for easier selection.

It was thus decided that the list would be limited to those who, during the study period, were working at the Public Service Department (PSD) only. Both the Management and Professional and the Supporting staff group 1 are known as the 'common services'. Personnel from these services are not attached permanently to any departments or ministries, but are routinely transferred from one department or ministry to another after a few years. The move can be lateral[1] or by virtue of promotion. The fact that these officers are routinely moved makes the selection of control group participants from one particular department, in this case the PSD, have minimal bias. They are typical and representative of all officers of the two managerial groups who are serving in any other departments or ministries.

The PSD has eleven different sections, but the main offices for all sections are located within the office complex at Putrajaya. Whether the control group has attended the same courses before or not is not important, as this research is measuring variation within the study period. What is clearly important is that during the time period the study group was attending courses and the control group was not.

For the study group, the selection of study participants was by **one-stage cluster sampling** [7]. Training programmes (in this case, **courses**) which were intended for the two managerial groups made up the clusters. Simple random sampling was then used to select some (n) clusters, all members of which are selected to be the study sample (Figure 3.2).

This sampling technique has several advantages to this study. One is that the administration of questionnaires became more manageable. For this study, the questionnaires were effectively distributed in the classrooms, when all participants were present. Briefings and explanations of the research were also done at the same time, and participants were verbally thanked before they even began to answer the questionnaires. The other significant advantage was not having the need to work with

[1] The officer taking up a new position which is of the same managerial level.

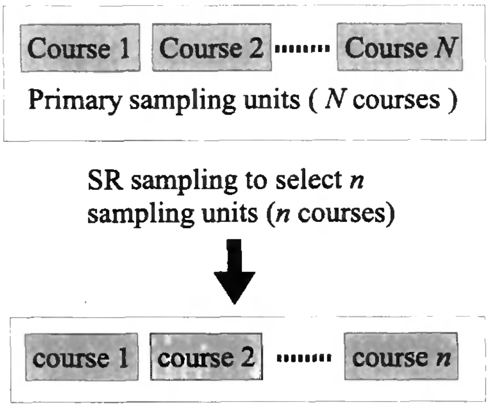

Course 1
Course 2
Course *N*
Primary sampling units (*N* courses)
SR sampling to select *n* sampling units (*n* courses)
course 1
course 2
course *n*

Figure 3.2: Sampling for the study group.

the list of all participants in the selected courses, because all of them were selected as respondents.

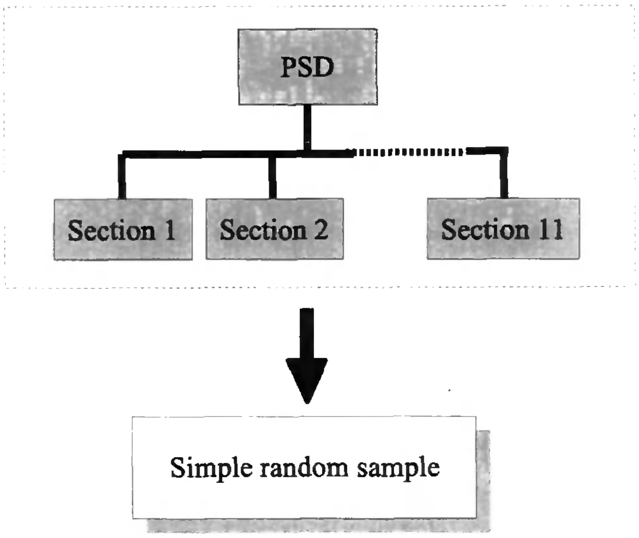


Figure 3.3: Sampling for the control group.

For the control group, the selection was by a straight-forward simple random sampling (Figure 3.3). The name list of all the Management and Professional officers and the Supporting staff from all sections within the PSD, minus those who were away from the office, was the sampling frame. Sections in the PSD were not considered in the sampling process because for this study, section is not a factor of interest.

3.4 Measurement Tools (Questionnaires)

This study mainly attempts to measure two factors: the *reaction* and the level of *learning* of the course participants. These latent factors are the first two levels of the Kirkpatrick's levels of evaluation. Reaction is measured by the Course Experience Questionnaire (CEQ), while Learning is measured by the newly developed Learning Questionnaire (LQ). Besides measuring these two main factors, this study also utilizes the General Health Questionnaire (GHQ), which measures the general health of the study participants as background information on their psychological health. In this section, two of the questionnaires, namely the CEQ and the GHQ are introduced. The LQ is only introduced in Chapter 4. It is given special attention because the LQ is the main measurement tool for this study.

3.4.1 The Course Experience Questionnaire (CEQ)

This survey instrument is being widely used in the Australian higher education system, and is starting to be used in some institutions in the UK [96]. Several researchers elsewhere had also used the CEQ, Espeland and Indrehus among others, who used it in their study of students satisfaction with nursing education in Norway in 2003 [24].

The CEQ was originally developed at Lancaster University, UK in the 1980s. A later version was developed and tested in Australian universities in 1989 [77]. Items are measured on 5-point Likert scales from 'Strongly disagree' to 'Strongly agree'. Some of the items are negatively worded, which thus needed re-alignment before analysis. This questionnaire is designed to measure differences in the quality of teaching between comparable academic organizational units in higher education systems [24]. Applied to this study, it could measure differences in the quality of training between or among the different levels of factors, namely: *gender, centre, ethnic, age, service sector, service group*, and *experience*.

The CEQ was developed based on the conclusion that there are real differences in teaching quality and that these variations could be measured [77]. The extensive development work of the CEQ took off with the testing of Version 1 in May 1989 in one Australian higher education institution. Based on three different questionnaires, it had 80 items which were later reduced to 57 as a result of the trial. The second version was tested on a different sample of 300 students in nine courses in two institutions. Internal consistency was reported to be reasonable, and discriminant validity was also reported with clear evidence [77]. Subsequently, version 2 was re-analysed to prepare for a national trial based on the recommendation of the Australian Higher Education Performance Indicators Research Project. The final instrument of the version 2 had five subscales, as presented below.

Good teaching Indicates good teaching practice of the trainers.

Clear goals Relates to the expectations and the objectives of the programmes.

Appropriate workload Relates to the amount of work on the part of the participants.

Appropriate assessment Indicates the agreement of the participants about how they are assessed and evaluated.

Emphasis on independence Relates to the amount of choice available to the participants regarding their works.

The author reported doing a series of factor analyses, which confirmed the scale structure of the instrument. Cronbach's alpha and item-scale total correlation showed convincing evidence of stability. Cronbach's coefficient alpha provides actual estimates of reliability [67]. A more detailed discussion on reliability is provided in section 3.7.1 on page 49.

Scale validity was evaluated by examining the strength of the relations between scale totals and three external criteria: quality of student learning, student satisfaction, and lecturers' reports of their own attitudes to teaching. In all three instances, findings were in accordance with the theoretical model and previous studies, supporting the instrument's validity. Overall, the author concluded that the CEQ possessed good psychometric qualities. Both the scale structure and its discriminatory power were not affected by any particular response categories, the wording, or the method of sampling. Self-selection of courses by students and averaging over several members of academic staff did not cause adverse effects either.

Current Form of the CEQ

Over the years the CEQ has gone through several developments. Currently, there are two versions of the form. The short form has 23 items (CEQ23), and the long form has 30 items (CEQ30), like the one used in the national trial in Australia in 1991. The short form is the most widely used. It consists of only four of the original scales, plus one new scale. **Generic Skills** took the place of **Emphasis on Independence**, because the latter was found to have comparatively weaker scale structure [96]. Wilson *et al.* also mentioned that at that time, there had been an increasing awareness of the need to produce graduates who possessed skills relevant to employability and lifelong learning. These skills which graduates and employers considered *generic to workplace competence* had been identified. The Generic Skills scale of the CEQ was then designed to assess the extent to which graduates perceived their courses as developing these skills.

Reliability and Validity of the CEQ

After the new Generic Skills scale was developed, Wilson, Lizzio and Ramsden did a study to investigate the validity and reliability of the new instrument [96]. The long form of the instrument (CEQ30) was combined with the new scale of six items producing CEQ36. The short form (CEQ23) consisted of the full version of the then new Generic Skills, plus a shortened version of the other four scales, excluding Emphasis on Independence. The study was conducted using three samples between 1992 and 1994. The Cronbach's alpha for the three samples showed moderate to high internal consistency for all scales. The short form showed slightly lower alpha on some scales, but nevertheless it was concluded that both forms were reliable instruments. In the short form all the items tend to load on distinct factors reflecting their assigned scales [17], but in the longer versions (CEQ30 and CEQ36) there is a consistent tendency for a few items on the Good Teaching scale and the original Emphasis on Independence scale to load on other factors [15,56].

A factor analysis of the items confirmed the original five structures and the new Generic Skills scale. All items in both the 1992 and 1994 samples loaded on one of the six factors, with only two items loaded significantly on more than one factor. Three items loaded on factors different to their nominated scale, prompting the author to suggest that students perceived the 'development of one's own academic interests' to be one of the generic competencies acquired through a university education. There were some cross-loading items, and these were the ones eliminated to create the short form CEQ23. The results of the short form were very similar. All 23 items loaded on their nominated scales. The factor structure of this shortened instrument was as stable as the long form, while having the advantage of cleaner relationships between items and scales [96].

Wilson *et al.* reported that the item confirmatory factor analysis showed moderate overall fit of the data to the model for CEQ36, and a good fit for the short form. The better fit of the short form was attributed to the absence of the Emphasis on Independence scale. For the long form, items of this scale showed low factor loadings and high structural coefficients for the error term.

Factor analysis at the scale level was carried out to examine whether item scores of the instrument could be aggregated to yield a single global score of teaching quality. By means of a higher order **path analysis**, three models were tested - a one-factor model (all scales), a two-factor model (Appropriate Workload and all other scales), and a three-factor model (Appropriate Workload, Generic Scales and all other scales). The results showed that the most suitable representation of higher order structure of the CEQ is the two-factor model; with Appropriate Workload as one factor and Good Teaching, Clear Goals and Standards, Generic Skills and Appropriate Assessment as the other factor (Figure 3.4). However, this result is not clear-cut and subject to debate [79].

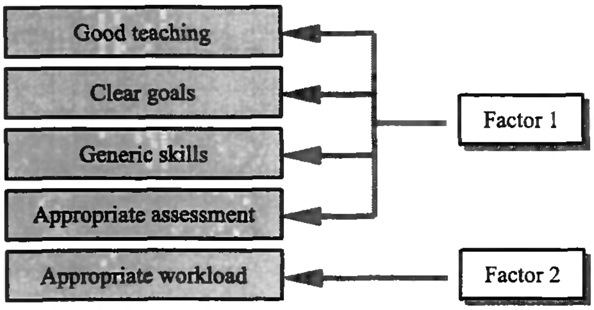


Figure 3.4: Higher order structure of CEQ36 and CEQ23.

Correlational analyses were conducted between the CEQ scale scores and a number of key external criteria. The criteria included approaches to learning and course outcomes (students' overall satisfaction, generic skills development, and academic achievement). From the results the authors concluded that the CEQ was clearly measuring aspects of the teaching environment which were systematically associated with students' reported learning processes. Furthermore, positive correlations with the learning outcomes was seen as strengthening the instrument's validity as a measure of teaching quality. Testing the instrument between two distinct fields of study, the authors further concluded on the discriminant validity of the scale. The two fields chosen were medicine and psychology, where programmes have distinct course objectives and teaching philosophies.

3.4.2 The General Health Questionnaire (GHQ)

The General Health Questionnaire (GHQ) was designed as an instrument to screen common mental disorder [93], to identify and measure psychological problem [18], or to detect non-psychotic psychiatric disorder [75]. In this study, the GHQ is used to provide background information on the general psychological health of the training participants. This information will then be examined for possible association with the score of 'learning'. This analysis could provide an indication whether the performance of the participants is related to their mental health.

The short version of the GHQ with twelve items and one factor is the most widely used, especially as a screening instrument [18]. The original factor structure of the GHQ is unidimensional. However it is not very stable as some studies in different settings had shown two and three factor solutions and multiple cross-loadings [93]. Though the number of factors yielded varies between studies, factor names have been quite common, such as symptom factors for anxiety and depression, factors related to coping, feelings of incompetences, somatic complaints, sleeping difficulties, and social functioning [18]. This study examines the factor structure of the twelve item instrument (GHQ-12) within this setting, using item total as the score.

There are four possible methods of scoring the GHQ. The methods are as follows:

1. **GHQ scoring.** It uses the score of 0-0-1-1 for all items.

2. **Likert scoring.** This method uses the more straight-forward scoring of 0-1-2-3 or 1-2-3-4 for all items.

3. **Modified Likert scoring.** The scoring is 0-0-1-2.

4. **C-GHQ.** 0-0-1-1 is used for positive items, and 0-1-1-1 is used for negative items.

In this study the Likert scoring of 1-2-3-4 is used. Since the sum of the twelve items is taken as the GHQ score, the score ranges from 12 to 48. For this type of scoring, a typical score is between 23 and 24. Any score above 27 is an indication of distress, and anything above 32 suggests severe problems [31].

3.5 Administration and Data Collection

The collection of data of this study involved approaching public service officers during their working hours, and course participants at INTAN in their classroom hours. Because of that, formal notifications were sent to the relevant heads of divisions to inform them and to get their approval. First, application to do this research was sent to the Director of INTAN. He was to be the 'champion' of the research at INTAN, where the main data was to be collected. For the control data which was to be collected from other divisions of the Public Service Department, an application was sent to the Director of the Human Resource and Management Services Division.

An email sent to the Director of INTAN was replied by the Head of the Corporate Unit, Mr. Razali Malek on behalf of the Director of INTAN, who had given the green light for the study and the collection of data. Following that, several emails were exchanged with the Registrar of INTAN, Mr Hadzir Md Zain, discussing the logistics of the study. With regards the Public Service Department (PSD), an email was sent directly to Mr Yasin Salleh, the then Director of the Human Resource and Management Services Division. He promptly replied the email giving his permission for data collection, at the same time forwarding a copy to Ms Munirah A. Bajanuddin, a Deputy Director, with whom further correspondence would be made. Ms

Bajanuddin also helped with the arrangement of the distribution and collection of the questionaires. Unfortunately, none of the emails used in correspondence with both INTAN and PSD was kept as a record, thus is not available to be included in this thesis.

Data needed for this study were obtained through the use of three questionaires, namely the GHQ, the CEQ, and the LQ. For the study group, questionnaires were distributed in the classrooms, where the participants attended the courses. For the control group, the distribution of questionnaires was done through the administrative officer of each division. Completion of the questionnaire was done in the respondents' own offices. The LQ consists of two sets, the pretest and the posttest questionnaires. For the study group, the pretest questionnaires were distributed on the first day of the course. It was important to arrange for the participants to complete and submit the questionnaire there and then. This arrangement helped to minimise non-response, as well as to control for the bias as a result of completing the questionnaire at leisure.

Similarly, the posttest questionnaires were distributed on the last day of each course. A specific time was allocated for the participants to complete the questionnaire, shortly before they were dismissed from the courses.

3.6 Data Preparation and Screening

Data screening in structural equation modeling (SEM) is crucial because most widely used estimation methods in this analysis require certain assumptions about the distribution of the data [54], which is **multivariate normality**. Kline (2005) also points out that data related problems may contribute to the failure to obtain a logical solution, and this could be mistaken for model faults. Practically, inference about means based on the assumption of multivariate normality is unlikely to be misleading, so long as the distributions are not obviously skewed or long-tailed, and the number of samples are reasonably large [57].

The other concern is cases of outliers. A univariate outlier is when a case has an extreme score on a single variable. We will take 'extreme' to mean the score is more than three standard deviation (3σ) distance from the mean. A multivariate outlier can occur in two ways. One way is when scores are extreme on two or more variables. Secondly, scores may be fashioned in an atypical pattern, such as positioned between two and three standard deviations on all variables. Not being beyond three standard deviations, it is not a case of a univariate outlier, but the pattern is not typical in the sample.

Missing Data

There are two main types of **missing data** or missing observations: *systematic missing data*, and *ignorable missing data.* Most methods suggested to deal with missing data assume that the missing data are of the second type. Ignorable missing data is called *missing at random* (MAR) if the observations are missing by chance. If on top of that, it can be assumed that the missing is unrelated to any other variables, then it is known as *missing completely at random* (MCAR).

In either case, missing data can cause bias in the analysis. Since bias from missing data depends on the proportion of the missing data and not on the number of observations, it cannot be reduced by increasing the sample size (Fayers *et. al*, 2001). Kline [54] describes a few general categories of methods for dealing with missing observations as follows.

1. *Available case methods.* This method assumes MCAR situations. Cases are deleted in either of the two ways:

- In **listwise deletion**, cases with missing scores on any variables are excluded from all analyses. If the number of missing observations are large, then the effective sample size is substantially smaller than the original sample. The advantage is that all analyses are conducted using the same number of cases. Fayers and Machin (2001) [27] call this approach as complete-case analysis. They point out that having a reduced sample may produce misleading results. Therefore they do not recommend this approach unless the proportion of cases with missing scores is less than 5 percent.

- In **pairwise deletion** (available-case analysis-Fayers *et. al*), cases are excluded only if they have missing data on variables involved in a particular computation. Consequently the sample size varies from analysis to analysis. This is a drawback for SEM and other multivariate methods, because it may produce a 'nonpositive definite' variance matrix. Because of this, pairwise deletion is not recommended for SEM unless the number of missing observations is small. In this study, this approach is used in analyses other than SEM.

2. *Single imputation methods.* There are four techniques used under this heading:

- **Mean substitution** is the most basic, where the missing score is replaced with the overall sample average. This technique however, tends to distort the underlying distribution, reduce variability and make the distribution more peaked at the mean. To correct for the reduced overall standard deviation, Fayers *et. al* suggests multiplying the new standard deviation with

$$f = (\frac{N-1}{N-M-1})^{\frac{1}{2}},$$

where N is the total number of respondents and M is the number of missing respondents.

- **Regression-based imputation** uses the predicted value to replace the missing score. The value is generated using multiple regression based on nonmissing scores on other variables.

- In **pattern matching**, the replacement is done with a score from another case with similar profile on other variables.

- In **random hot-deck imputation**, missing scores are replaced with those on the same variable from the nearest complete record.

3.7 Analyses

In this study there are five main types of analysis, which are as follows:

1. **Reliability of the scales.** This is examined mostly by Cronbach's alpha, and by an alternative measure which we call alpha*. This topic is introduced further in Section 3.7.1.

2. **Confirmatory factor analysis (CFA).** Structural equation modeling (SEM) is used as the CFA procedure, which tests whether the datasets fit their hypothesized models. The AMOS 6.0 [4] computer software is used for this purpose. The SEM is introduced in greater details in Section 3.8. The results of the analysis is discussed in Chapter 6.

3. **Principal variables analysis (PVA).** This analysis explores the reduction of the dimensions of the datasets. This is achieved by selecting a few of the variables that contribute to the majority of the overall variance. Chapter 7 further introduces this analysis and discusses the results.

4. **Tests of differences.** Analysis of variance (ANOVA) and t-test are utilized to compare the scores of the three questionnaires and their subscales, between or among the different levels of seven demographic factors. The factors are *gender*, *ethnic origin*, *age*, *centre*, *service sector*, *service group*, and *experience*. The results of the tests are discussed in Chapter 8.

5. **Analysis of covariance (ANCOVA).** This analysis is only carried out on the Learning data. The pretest Learning is used as the covariate, and the posttest Learning is the dependent variable, in examining the effects of the seven demographic factors. Further discussion on the this topic is presented in Section 9.4, along with the results of the tests.

Most of the analyses are carried out using R software [88], except for SEM which uses AMOS 6.0 [4]. The softwares are run on a Twinhead E12B notebook [21], running on Intel(R) Pentium(R) M (1500 MHz processor) with Microsoft XP Home Edition.

3.7.1 Reliability and Cronbach's alpha

A valid measuring instrument measures what it purports to measure in the context in which it is to be applied (Nunnally and Bernstein) [67]. Three aspects of validity are (i) *construct* validity, (ii) *predictive* validity, and (iii) *content* validity. The first measures psychological attributes, the second establishes statistical relationship with a particular criterion, and the third samples from a pool of required content. The first two are validated by correlations among various measures, but the third validity is usually based on opinions. All three have much in common, but they also have important differences.

Reliability has two definitions. The first is internal consistency, which is always desirable. The second definition of reliability is stability over time. Assessment of the second definition is usually by test-retest or parallel forms.

According to Nunnally et. al., a **reliability index** (r_{it}) is the correlation between a set of scores on a given test (x_1) and the corresponding true scores. The correlation of variable x_1 with the sum of the k variables approaches the correlation of variable x_1 with true scores (the sum or average of scores on all possible variables) as k approaches infinity. This in turn is equal to the square root of the average correlation between all pairs of tests in the domain, and this relationship is shown in (3.1).

$$r_{1(1...k)} = r_{1t} = \sqrt{\bar{r}_{1j}} \tag{3.1}$$

The $\bar{r}_{ij}$ may be estimated by the reliability coefficient for test x_1. The **reliability coefficient** is the ratio of the variance of true scores to the variance of observed scores. Cronbach's coefficient alpha (α) estimates a reliability coefficient using item intercorrelations.

High reliability is always sought after but the standard of acceptable reliability depends on the type of decision to be made (Nunnally *et al.* [67] p. 249). Tests used to contrast groups need not be as reliable as tests used to make decisions about individuals. Further, Nunnally *et. al.* explained that test validity is not always limited by limited reliability. A relatively valid but somewhat unreliable test should not be replaced by a less valid test.

The reliability of test scores is often evaluated using correlations of items with the true score and with each other. Nunnally *et al.* showed that this type of score increases with the number of items. In an example of 20 items which had an average correlation of 0.25, the expected correlation of an item on the true score was 0.87.

Cronbach's alpha is the expected correlation of one test with another test of the same length and measuring the same thing. It can also be viewed as the expected correlation between an actual test and hypothetical form [67]. An illustration (taken from Nunnally *et al.*) as to how alpha is calculated from a variance-covariance matrix is presented in the tables below.

Var/Covariance

	x1	x2	x3
x1	10	5	4
x2	5	5	3
x3	4	3	8

Correlation

x1	x2	x3
1	0.71	0.45
0.71	1	0.47
0.45	0.47	1

From the variance-covariance matrix, the correlation matrix is calculated by dividing each off-diagonal term (covariance) by the square root of the product of the on-diagonal terms (variances) that appear in the same row and in the same column (Nunnally et al. page 165). The sum of all elements in the covariance matrix is $\overline{R} = 6.26$. Using [3.2] (with k = 3) gives the value of alpha equals 0.78.

$$r_{kk} = \alpha = \frac{k}{k-1}\,\frac{\overline{R}-k}{\overline{R}} \tag{3.2}$$

The fact that this value is higher than the average correlation ($\overline{r}$ =0.54) is then highlighted, implying that the average correlation between the items tends to underestimate the reliability of composite measures.

Apart from that it is also mentioned that standardization which takes place during the calculation has no essential effect on the value of reliability. However it might not hold true for small k, but it does when k is large. Nunnally *et al.* also suggest that Cronbach's alpha be used with other estimates of reliability. It sets the upper limit for the tests, and when it is low, there is no point in doing the other tests. They also acknowledge that it ignores certain potentially important sources of measurement error, but the difference it makes is negligible.

Besides Cronbach's alpha, this study also uses an alternative reliability index which we call **alpha***. Alpha* measures implied reliability had there been only two items in the scale. A large difference between the values of the Cronbach's alpha and the alpha* indicates that the scale has a lot of items to compensate for weak correlation between the items and the latent variable.

Cronbach's alpha which is used to estimate the reliability of a k-item scale is also defined as Equation (3.3) (Nunnally et. al. pp. 234).

$$\alpha = \frac{k}{k-1}(1 - \frac{\Sigma\sigma_i^2}{\sigma_y^2}) \tag{3.3}$$

where $\Sigma\sigma_i^2$ is the sum of variances, and σ_y^2 is the variance in total scores. The part in parenthesis in (3.3) could also be presented as (3.4). If, assuming to some order of approximation that $Var(x_i) = \sigma^2$ and $Cov(x_i, x_j) = \rho\sigma^2$ then (3.3) becomes (3.5), as presented in Nunnally et. al. (p 232).

$$1 - \frac{\sum_{i=1}^{k} Var(x_i)}{\sum_{i=1}^{k}\sum_{j=1}^{k} Cov(x_i, x_j)} \tag{3.4}$$

$$\alpha = \frac{k\rho}{1 + (k-1)\rho} \tag{3.5}$$

Rho and standardised alpha.

In 3.5, we can see that for $0 \leq \rho \leq 1$, and $0 \leq \alpha \leq 1$, the value of alpha approaches the value of rho ($\alpha \longrightarrow \rho$) as the number of items increases ($k \longrightarrow \infty$). The value of ρ is an estimation of the mean of inter-item correlation. From (3.5), ρ can be presented as (3.6).

$$\rho = \frac{\alpha}{\alpha + k(1-\alpha)} \tag{3.6}$$

The following inequality of 3.7 shows that for two scales with equal ρ, the one with more items has greater alpha. In other words, the more items a scale has, the more reliable it is, even though the average inter-item correlation remains the same.

$$\alpha^{(k+1)} = \frac{(k+1)\rho}{1 + (k)\rho} \quad > \quad \frac{k\rho}{1 + (k-1)\rho} = \alpha^{k} \tag{3.7}$$

In a scale, k is the number of items that make up the scale. If the scale had only two items, then (3.5) becomes (3.8), which is how alpha* is defined.

$$\begin{aligned}\alpha^{*(2)} &= \frac{2\rho}{1+\rho} \\ &= \frac{2\alpha}{(2-k)\alpha+k}\end{aligned} \tag{3.8}$$

Thus alpha* (3.8) is the value of alpha (α) implied if the apparent correlation ρ held for a two item scale. In other words, it shows the value of implied reliability had there been only two items in the scale. Values of this alpha are then used as a standardised index to compare reliability among the scales used in this study which have different numbers of items.

3.8 The Structural Equation Modeling (SEM)

Structural Equation Modeling (SEM) is given a special attention in this thesis because of its important contribution to the development of the Learning Questionnaire (LQ). SEM is a general statistical modeling technique, consisting of a collection of techniques that allow examination of a set of relationships between one or more independent variables, and one or more dependent variables. Both the dependent and the independent variables can either be continuous or discrete [90].

In SEM, input to the analysis is the sample covariance matrix. A model then produces estimated population parameters, which are combined to create the estimated population covariance matrix. This population covariance matrix is then compared with the sample covariance matrix (Figure 3.5). If the difference between them is small, the model is said to fit the data well.

Unlike ordinary regression analysis, SEM considers several equations simultaneously. The same variable can be a predictor in one equation and a criterion in another. SEM presents this system of equations in a structural model and measurement models, which is often presented graphically to aid viewing. The two models represent two main steps in structural equation modeling: (1) validating the measurement model, and (2) fitting the structural model.

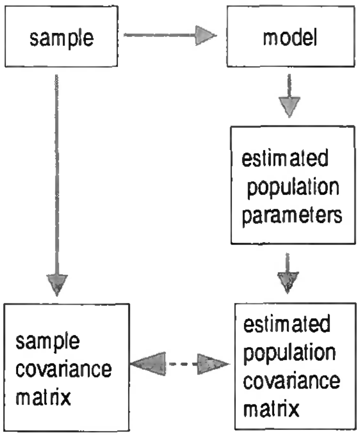

sample
model
estimated population parameters
sample covariance matrix
estimated population covariance matrix

Figure 3.5: SEM model.

3.8.1 Structural and Measurement Models

A **structural model** summarizes the relationships between latent variables. A latent variable is unobserved or unmeasured variable which corresponds to a concept, thus it is hypothetical [12]. Bollen also calls this model the **latent variable model**. To introduce the notation for a structural model, we use the relationship of the course experience (of the CEQ scale), to its five factors. The five factors are (i) good teaching (GT), (ii) clear goals (CG), (iii) generic skills (GS), (iv) appropriate assessment (AA), and (v) appropriate workload (AW). The structural model of the CEQ which is first presented as Figure 3.4 on page 43 is presented again as Figure 3.6.

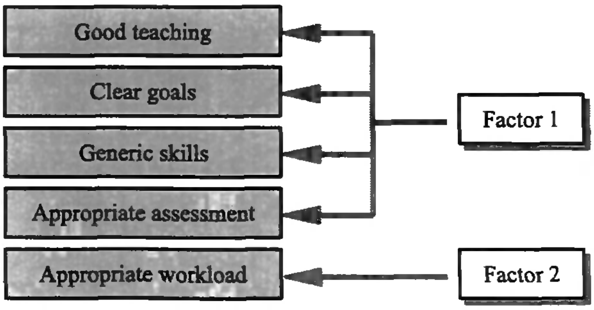


Figure 3.6: Structural models of the CEQ.

Figure 3.6 shows two structural models, each related to an unnamed factor. In the first structural model, Factor 1 is explained by four of the latent CEQ subscales, while in the second structural model, Factor 2 is explained by one subscale. These two models represent a theory, a test of which is only possible if we collect observable measures of the subscales, which themselves are latent factors. Developers of the CEQ have shown that the subscales or the factors have between three to six indicators. Relationship between each latent factor and its indicators represents a **measurement model**. In order for the structural model to be measured, the measurement models have to be validated first.

The task of fitting the structural models is primarily accomplished through path analysis with latent variables [33]. Covariances predicted by the model are compared to the actual covariances in the data. Effect sizes ('regression weights' in the AMOS software) are **structural** or **path coefficients** estimated by the computer program. As is most commonly done, maximum likelihood estimation (MLE) is used to estimate the coefficients. It assumes that samples are large and that indicator variables have multivariate normal distribution. Furthermore, it also assumes valid specification of the model and continuous indicator variables.

In the SEM, error terms are explicitly modelled, making path coefficients unbiased. This is in contrast to regression analysis where coefficients are made less effective by measurement error. However, if the error is high, the estimates of the SEM path coefficients will be less reliable and less trustworthy as well [33].

Normally, coefficients in SEM are standardized. Interpretation of these is not much different to interpreting standardized regression coefficients, where they give the relative importance of each independent variable. If the standardized structural coefficient is 2, it means that an increase of 1 unit in the independent corresponds to an increase of 2 units in the dependent latent variable. In the model, the path is significant at 0.05 level if the Critical Ratio (CR) is $>$ 1.96. Similarly, estimated covariances among the latent variables are significant if CR $>$ 1.96. However, unstandardized coefficients are preferable when comparing across groups. This is because across groups, indicators may have different variances, as may latent variables, measurement error terms, and disturbance terms.

3.8.2 Degrees of Freedom

The degrees of freedom in SEM are the difference between the amount of unique information in the sample variance/covariance matrix and the number of parameters in the model to be estimated [90], or the difference between the number of correlations or covariances and the actual number of coefficients in the proposed model [45].

A simpler way of calculating the degrees of freedom is presented by Ullman (2001) and Stevens (1996). This approach is especially appropriate for cases of larger models where it is not easy to determine the number of variances and covariances. The number of data points or number of unique values in a covariance matrix is determined by

$$\frac{p(p+1)}{2}$$

where p is the number of measured variables. This is equal to the number of covariance elements below the diagonal plus the number of variance elements (diagonals). The difference between this value and the number of parameters is the degree of freedom used. The number of parameters to be estimated in the model consists of (i) variances of all independent variables, and (ii) regression coefficients.

Another way of calculating the degrees of freedom is the following (Hair, Anderson, Tatham and Black, 1998);

$$df = \frac{1}{2}[(p+q)(p+q+1)] - t$$

where : p = the number of endogenous indicators, q = the number of exogenous indicators, and t = the number of estimated coefficients in the model. Since degrees of freedom in SEM is calculated based on the data matrix, it is not affected by sample size like in other multivariate methods.

3.8.3 Model Identification

The population covariance matrix can only be estimated from an *identified* model. An identified model is one which has unique numerical solution for each of the parameters. Ullman (2001, [90]) suggests the following guidelines to check whether a model can be identified.

1. The first step is to count the numbers of data points and the number of parameters to be estimated. This condition for identification is also known as **order condition** [45]. Data points are also known as some other names; Stevens (1996) refers to them as pieces of information, and Hair *et al.* calls them unique information. Kline [54] simply refers to them as the number of observations.

   If there are more data points (pieces of information/unique information) than parameters, the model is overidentified[2]. If they are equal, the model is just identified. Otherwise if data points are less than parameters, the model is underidentified and parameters cannot be estimated.

   In just identified models, estimated parameters perfectly reproduce the sample covariance matrix. In this case, only paths in the model can be tested, but adequacy of the model cannot. Underidentified models need to be adjusted to become just identified or overidentified before they can be tested. This is done by either fixing, constraining, or deleting some of the parameters.

2. The second step in model identification is looking at the measurement portion of the model. There are two parts to this. Part one is establishing the scale of the factor. This is done either by fixing the variance of the factor to 1, or fixing to 1 the regression coefficient from the factor to one of the measured variables (the one chosen is called **reference variable**). The latter option gives the factor the same variance as the measured variable. If the factor is an independent variable, either option is acceptable. If it is a dependent variable, most researchers use the second option [90].

   Choosing which indicators to be the reference variable is arbitrary, otherwise it makes sense to select the one with the most reliable scores [54]. Fixing the variance of the factor to 1 makes the factor standardized, similar to standardizing variables by transforming them to z-score. Otherwise, fixing one of the

[2] A necessary condition for analysis.

regression coefficients to 1 keeps the factors unstandardized. The latter option is known as **unit loading identification** (ULI) constraint, while the former as **unit variance identification** (UVI) constraint. UVI specification also implies that the loadings of all indicators for the factor can be freely estimated with sample data. In confirmatory factor analysis (CFA) however, setting the variance to 1 is the more common way of assigning a scale [85]. Once factors are scaled (through ULI or UVI) the total number of free parameters is reduced by one for each factor.

The number of factors and the number of measured variables loading on each factor will determine the identifiability of the measurement portion of the model. If there is only one factor, the model may be identified if the factor has at least three indicators with nonzero loading and the errors are uncorrelated with each other.

In models with two or more factors, the number of indicators is again considered. If each factor has three or more indicators, the model may be identified if (i) errors are not correlated, (ii) each indicator loads on only one factor, and (iii) the factors are allowed to covary. If there are only two indicators to a factor, the model may be identified if (i) errors are not correlated, (ii) each indicator loads only on one factor, and (iii) all variances or covariances among factors are not zero.

3. Examining the structural portion of the model is the third step. If none of the dependent variables predicts each other then the structural portion of the model may be identified. Otherwise, the dependent variables need to be recursive[3] for the structural portion to be identifiable.

   Non-recursive models need further two condition for identifiability. Each equation in the model (separately) needs to have at least the number of latent dependent variables - 1 excluded from it. Secondly, the information matrix is full rank and can be inverted.

[3] No feedback loops, (ie. two arrows with different direction connecting two dependent variables) and no correlated disturbances among them.

3.8.4 Model Evaluation

After the model has been specified and the parameters estimated, the next step is asking the major question of whether the model is good. A well fitting model should have only small and non-significant differences between the sample covariance matrix and the estimated population covariance matrix. One way of testing this goodness of fit is a chi square test, evaluated at the model's degrees of freedom. A non-significant χ^2 value indicates a good fit.

A very rough rule of thumb for indicating a good-fitting model is that the ratio of χ^2 to the degree of freedom is less than 2 ([90]). Some researchers use 3 for 'reasonably' fitting, and 1 for good-fit (Trusty, Ng and Watts [89], following Arbuckle & Wothke [4]).

Computer softwares give two chi-squares: the **independence model chi-square**[4] and the **model chi-square**. Independence model chi-square tests the hypothesis that there is no relationship among variables. Since there should be some relationships, the test should be significant. Model chi-square is the one which should be non-significant, to indicate model fit.

The problem with chi-square is that its power depends on sample size, just like many statistical tests for model fits. With large samples, a small difference may cause the statistics to be significant [41]. In small samples, the computed χ^2 may not be distributed as chi-square. Sometimes, when the assumptions underlying a chi-square test are violated, the probability levels are inaccurate [90].

Because of these reasons, there are many other fit indices which have been developed to indicate model fit. Some of these indices are considered because of their simplicity [41], as well as being less dependent on the size of the sample. Computer software AMOS 6.0 which is used in this study produces goodness of fit tests as in Table 3.1, presented together with values of good-fit thresholds.

The normed fit index (NFI), comparative fit index (CFI), and the root mean square error of approximation (RMSEA) are all based on comparing the χ^2 value to the χ^2 value of other models. The NFI compares the value to the value of χ^2 for

[4] Also known as 'Null model chi-square'.

Table 3.1: Goodness-of-fit tests with good-fit indications.

Index	Good fit indication
CMIN/DF	< 2
Goodness of fit (GFI)	Close to 1
Adjusted goodness of fit (AGFI)	Close to 1
Parsimony goodness of fit (PGFI)	Close to 1
Non-normed fit index (NNFI or TLI)	Close to 1
Normed fit index (NFI)	> 0.90
Ratio of χ^2/df	< 2.0
Comparative fit index (CFI)	> 0.95
Root mean square residual (RMR)	Close to 0
Root mean square error (RMSEA)	< 0.05

the independence model. One drawback is that it might underestimate the fit of the model in well-fitting models with small samples [90]. The CFI also compares χ^2 values but with a different approach. The RMSEA meanwhile does the comparison with a perfect (saturated) model. This index is also less preferable with a small sample.

The goodness-of-fit (GFI) and the adjusted goodness-of-fit (AGFI) indices indicate the proportion of variance in the sample accounted for by the estimated population covariance matrix. The GFI has been suggested as analogous to R^2 in multiple regression. The AGFI adjusts the GFI value for the number of parameters. The parsimony GFI is developed from the GFI to take into account a parsimony adjustment. The root mean square residual (RMR) index shows the average difference between the the sample variances and covariances, and the estimated population variances and covariances. A small value is consistent with a good fit.

3.8.5 Modification Indices (MI)

It is very often in SEM that the initial model does not fit the data well [54]. The task following the initial analysis thus is to try improve model fit by model modification or respecification.

AMOS 6.0 also produces **modification indices** (MI) as output. Each index of the MI reflects the predicted decrease in χ^2 value if a single fixed parameter or equality constraint is removed from the model and the model is re-estimated.

An estimate of how much the coefficient would change is also presented in the MI output.

For MI that relates to the covariances, it has to do with the decrease in χ^2 if the two error terms are allowed to correlate. In the case of regression weights, the MI has to do with the decrease in χ^2 if single-headed arrows are added to the path.

3.9 Application of SEM in this Study

In this study, SEM is applied to all three questionnaires: the GHQ, the CEQ and the LQ. The process consists of two main steps: (1) validating the measurement model, and (2) fitting the structural model. The GHQ is a single factor model, thus it only has one measurement model. The CEQ has five measurement models, and two proposed structural models. The LQ has three models at the measurement level and one proposed structural model.

Validating the Measurement Model

The GHQ is a single factor model indicated by twelve indicator items. Validating the measurement model of the GHQ means testing whether all twelve indicators load on this factor.

The measurement model of the CEQ is based on the original development literature of the CEQ23 questionnaire (Wilson, Lizzio and Ramsden, 1997) [96]. The concept of course experience is measured by five factors or latent variables which we call the CEQ subscales. Figure 3.7 indicates the five subscales and the number of indicators for each. In the proposed LQ, there are three Learning factors, each of which is indicated by ten items. The LQ is discussed in futher details in Chapter 4.

Fitting the Structural Model

In this study, there are two structural models to be tested: one is of the reaction model which is measured by the CEQ, and the second one is of the Learning which is measured by the LQ. The GHQ does not have a structural model because it has only one latent factor. The structural model for reaction is as in figure 3.4 on page 43, while that for Learning is as in Figure 4.1 on page 64.

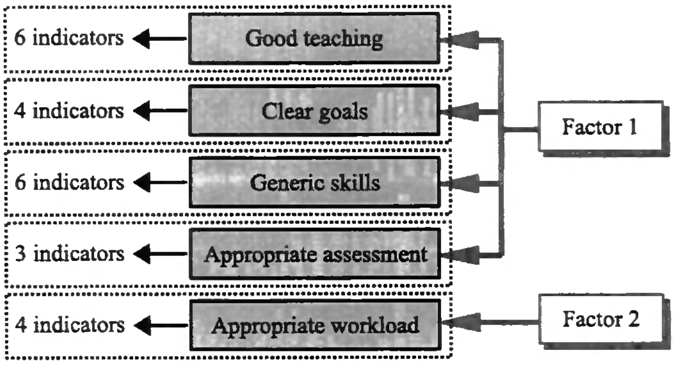


Figure 3.7: Five measurement models of the CEQ.

The main objective of model fit analysis is to test whether a specified model fits the data. For the model of reaction, we are interested to see whether the indicator items load on their intended factors. In one of the two higher order models, we also would like to test whether scores from four out of the five variables (excluding the AW subscale) load on one factor, while the rest of the indicator items load on the AW subscale as proposed and presented in Figure 3.7. For the Learning model, the main interest is on the validity of the newly developed questionnaire. This would initially be indicated by indicator items that load on their intended factors.

Chapter 4

The Learning Questionnaire (LQ)

In this chapter we introduce the Learning Questionnaire (LQ), the main measurement tool in this survey. This questionnaire is specially developed for the purpose of measuring 'learning', in the context of this study. First we will take a look at why this questionnaire is necessary. Then we will discuss about the Learning Model as the basis of the LQ. Following that we will discuss the development and then the evaluation of the LQ.

4.1 The Need for the LQ

Training effectiveness has been most widely explained using Kirkpatrick's four-level model. The first level, *reaction*, relates to the enjoyment of training, perceived usefulness and difficulty, and general liking towards the programme. The second level of the model is *learning*, which is about getting new knowledge or new skills. The third level, *behaviour* change, is associated with changes in on-the-job behaviour. *Results*, which is the fourth level, is measured by examining the impact of training intervention to the firm or the organization.

In this study, the focus is only on the first two levels: *reaction* and *learning*. Learning is the main construct, and it is measured among all participants in the study group, as well as among all members of the control group. *Reaction*, on the other hand, is only measured among the study group. The justification for measuring *reaction* and *learning* has been discussed in Section 1.2 on page 3.

As this study involves a control group which is not attending any training during the data collection period, a measure of Learning specific to a particular course is not appropriate. Instead, the measurement used has to be general enough to be relevant to the control group as well, while being sensitive enough to measure the changes brought about by the training programmes attended by respondents from the study group. Furthermore, training programmes at INTAN vary widely in terms of subject areas, length of programmes, managerial levels of participants, approach of training, and other matters. A measure specific to a particular training programme would definitely be unsuitable to be used in INTAN.

Literature review failed to find a suitable questionnaire that could be used for measuring Learning in the context of this study. A questionnaire had to be developed specifically for this purpose. Having a tool that can measure Learning would improve the evaluation of training effectiveness at INTAN because the management would get indications as to how much participants have actually 'learned' from the training programmes. At the moment the current evaluation practice is restricted more or less to the *reaction* level only. In this study, a Learning Questionnaire (LQ) is developed and its value is investigated. The questionnaire is developed based on the Learning Model.

4.2 The Learning Model

The Learning Model is developed as an attempt to measure Learning. Referring to the Kirkpatrick's level of training evaluation, the concept of 'learning' is the second level. Figure 4.1 shows the structural model, where it is hypothesised that Learning affects three latent factors.

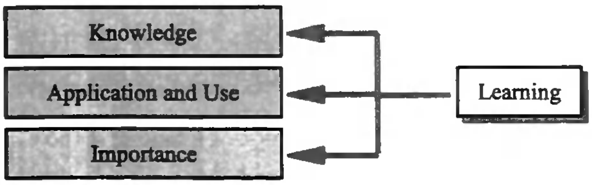


Figure 4.1: Structural Model of Learning.

In this proposed model, Learning is a construct consisting of three different subscales. The subscales are (i) perceived level of knowledge on the subject areas, (ii) application and use of the subject areas, and (iii) participants' attitude towards knowledge, which are named in the model as knowledge, application, and importance respectively. This is partly based on the works of Johnston, Leung, Fielding, Tin and Ho (2003), whose development work on their questionnaire yielded four factors: *(i) Future Use, (ii) Attitude, (iii) Knowledge*, and *(iv) Application and Use* [42].

For the Learning model, only three of the factors from Johnston *et al.* are used: *Knowledge*, *Attitude* towards the knowledge, and *Application and Use* of the knowledge in day-to-day tasks at the participants workplaces. The *Future Use* factor is not included because it is not relevant with the context of this study. *Attitude* is used and re-named as *Importance*. These factors are also chosen because they are in agreement with INTAN's concept of training, which is to have an impact on the levels of skills and knowledge, as well as on the attitude of participants.

The scope of INTAN's training covers eight different subject areas. Almost all training programmes of INTAN are developed within the context of at least one of these areas. The areas are:

1. Economic Management;
2. Financial Management;
3. Information Technology and Communication;
4. Human Resource and Organisation;
5. Social and Infrastructure Planning and Administration;
6. Land, Territorial, Regional and Local Government Administration;
7. International Relations and Foreign Affairs; and
8. Defense and National Security.

Besides these eight general subject areas, each course or training programme has its specific objectives. The objectives are usually stated as specific skills, knowledge or attitude, changes in the levels of which are intended as the target of the training programme. Thus in addition to the eight subject areas, two extra items are included in the Learning Model, namely (i) the *skills*, and (ii) the *knowledge*, targeted by the training programme. The eight subjects are called the *general areas*, while skills and knowledge are called the *focus areas.* The two focus areas are included in the model because we would like to know whether the training programmes have any impact on them. Thus we have ten areas altogether, and they will form the basics of the indicators of each of the Learning subscales.

The first Learning factor is *knowledge.* It refers to the participants' own perceived level of knowledge in the ten areas. Participants are asked to evaluate his or her knowledge in the subject areas and in the focus areas. The second factor, *application* (application and use), refers to the perceived level of usage of each of the ten areas in the participants' normal working environment. It indicates 'how much' each of the subject areas is applied and used in day-to-day work. The third factor, *importance*, refers to the attitude of the participants towards the importance of learning and re-learning the ten subject areas.

Each of these three latent Learning factors is indicated by the same ten items, representing the ten subject areas. The items are labelled as A1 to A10, B1 to B10, and C1 to C10 for knowledge, application, and importance respectively, as indicated in Table 4.1.

Figure 4.2 shows the Learning Model, where the three factors are the indicators for Learning. This specification, which identifies a common direct cause of all three first order factors, implies that the associations between the three factors are spurious, thus they are not shown to covary in the model. It means that in the model there is no arrow that connects them. The model also indicates that each of the first-order factors has two direct causes. One is Learning, the higher-order factor, and the other one is disturbances, which represent all causes other than Learning.

In the Learning Model, the three first-order factors are endogenous, while Learning is exogenous. Being exogenous, causes of Learning are unknown and not represented in the model, but a symbol of variance is put next to it. Symbols of variances are also placed next to the disturbances of the first-order factors because disturbances are considered as unmeasured exogenous variables. Scales of the disturbances are also set to 1.0.

As indicators, each of the thirty observed variables (A1 to C10) has measurement

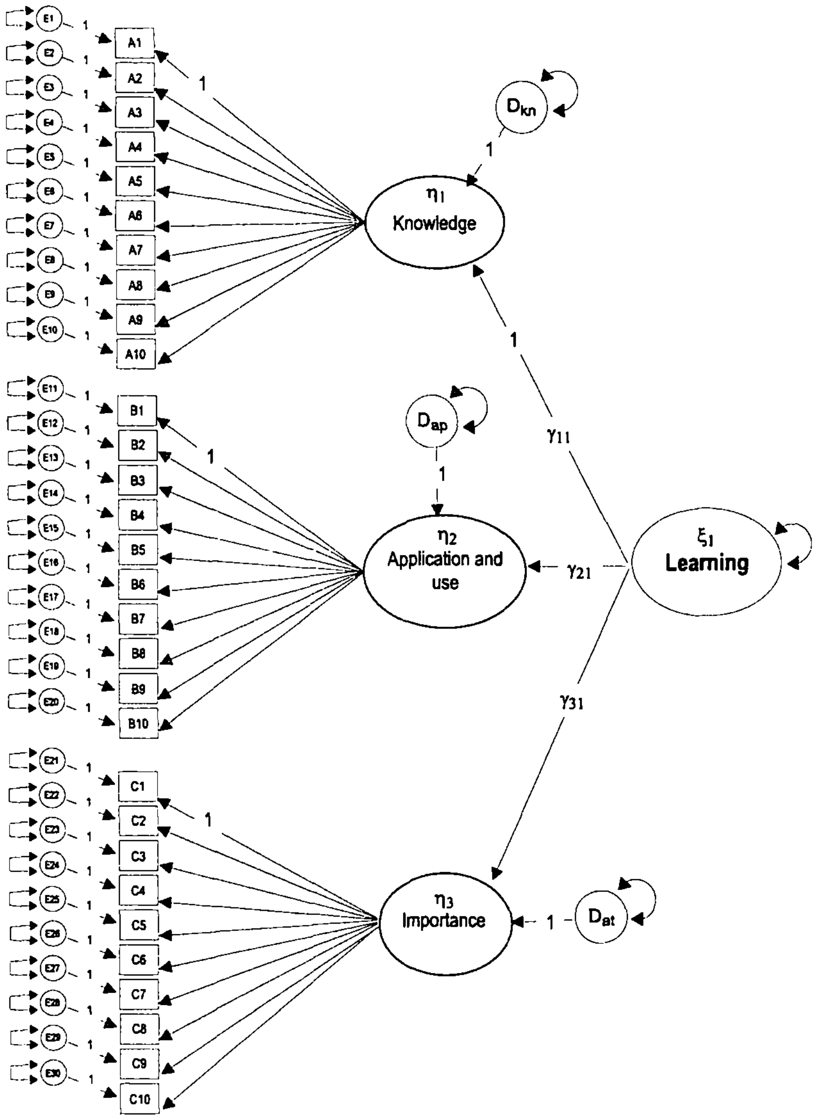


Figure 4.2: The Learning Model.

Table 4.1: Indicators of the Learning factors.

	Indicator items									
	General								*Focus*	
Factors	1	2	3	4	5	6	7	8	1	2
Knowledge	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Application	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
Importance	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10

Items of the LQ subscales and the subject areas they refer to.

Item	*Subject area*
1	Economic Management.
2	Financial Management.
3	Information Technology & Communication
4	Human Resource and Organisation.
5	Social & Infrastructure Planning and Administration.
6	Local Government Administration.
7	International Relations.
8	Defense/Security
9	*Knowledge* specific to the training programme.
10	*Skills* specific to the training programme.

error. Measurement errors are proxy variables for all sources of residual variation in their scores not explained by the three factors [54]. This is referred to as unique variance, which has two types: (i) random error in the indicators, and (ii) all systematic variance not due to the factors. Measurement errors are also unmeasured exogenous variables, therefore each of them has variance symbol next to it. To meet the conditions for identification, each of the measurement error is scaled to 1.0.

Having three first-order factors is the minimum requirement for the model to be identified [54]. Each of the first-order factor has 10 indicators, satisfying the requirement of at least two indicators. To set the scales of the first order factors (knowledge, application, and importance), one unstandardized loading for each is fixed to 1.0. The other possible option to set the scales is by fixing the variance of each of the factor to 1.0, effectively standardizing the factor. With the latter option, all thirty direct effects of first-order factors on the indicators would be free to be estimated.

For the second order factor Learning, one of its direct effects on the first order factors is fixed to 1.0, with the other alternative of fixing its variance to 1.0 instead. With the first option, only two of the effects of Learning on the first-order factors are free. If the alternative was used, all three effects would become free parameters.

4.2.1 Learning Model Specification

The general model for confirmatory factor analysis can be represented by either one of the following [12]:

$$x = \Lambda_x \xi + \delta \qquad \text{or} \tag{4.1}$$

$$y = \Lambda_y \eta + \epsilon \tag{4.2}$$

where y and x are observed variables, ξ and η are latent factors, and δ and ϵ are errors of measurement. For this discussion the second model is used. The notations of the Learning Model will be discussed in more details in Chapter 10, specifically in Section 10.1 on page 260.

In the Learning Model, there are three endogenous variables (η), namely *Knowledge* (η_1), *Application* (η_2) and *Importance* (η_3). These three are hypothesised to be affected by a single latent variable *Learning*, which is the only exogenous variable and represented by ξ_1. Each η_i is indicated by ten y_i's, the items of the questionnaire. The y_1 to y_8 refer to the eight subject areas, while y_9 and y_{10} refers to the specific skills and knowledge of a particular training programme.

It is hypothesised that the first ten indicators are linearly dependent on a single factor η_1 (Knowledge), the second ten indicators are linearly dependent on a single factor η_2 (Application), and likewise for the last ten on η_3 (Importance). None of these three factors influences each other. Each indicator y_i contains an error of measurement (ϵ_i) term which is assumed to be uncorrelated with the latent variables. Each of the first direct effect of the latent factors is fixed to 1.0 for identification of the model. These relationships of the measurement model are represented by the following matrix equation:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{30} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{2.1} & 0 & 0 \\ \vdots & 0 & 0 \\ \lambda_{10.1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{12.2} & 0 \\ 0 & \vdots & 0 \\ 0 & \lambda_{20.2} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{22.3} \\ 0 & 0 & \vdots \\ 0 & 0 & \lambda_{30.3} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \epsilon_{30} \end{bmatrix}$$

where $COV(\eta_i, \epsilon_j) = 0$ for all i and j, and $E(\epsilon_j) = 0$.

The appropriate structural model that relates *Learning* to the three factors (*Knowledge, Application,* and *Importance*) is as follows [12]:

$$\eta = \Gamma\xi + \zeta$$

and the relevant matrices as the following:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \gamma_{21} \\ \gamma_{31} \end{bmatrix} \begin{bmatrix} \xi_1 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{bmatrix}$$

The first element in Γ scales *Learning* (ξ_1) to η_1 (Knowledge). The remaining two factor loadings are free to be estimated. The variance of *Learning* is matrix ϕ_{11}, while another matrix, ψ, contains the variance of the first-order factors not explained by Learning.

4.3 Development of the LQ

The LQ is developed based on the Learning Model presented in Figure 4.2. The main objective here is to create a measurement tool that can measure 'learning'. This tool is intended to be used in INTAN, together with another tool that measures 'reaction', to help with answering the question whether training programmes at

INTAN are effective. Before that can happen, its value must be examined.

Based on the Learning Model, the latent factor Learning is indicated by three factors, namely knowledge, application, and importance. Each of these three Learning factors are in turn measured by ten indicators, which relate to ten subject areas. Of the ten subject areas, eight are called the **general areas** and two are called the **focus areas**.

In the LQ, the three factors are divided to a section each. Knowledge is measured in the first section, application is measured in the second section and importance in the third section of the questionnaire. In the first section, respondents are asked about their knowledge in each of the subject areas. As participants who attend a training programme work in many different departments or ministries, and doing distinctly different jobs, it is anticipated that many of them will have different levels of knowledge in the different subject areas.

In the second section they are asked about how much they think they use their knowledge in the ten subject areas at their workplaces. Some participants may need in-depth knowledge of a particular subject in order to do their job effectively, while the others might need some command of several different subject areas to be efficient. A senior officer who heads a division with several sub-divisions needs to have a good command of knowledge in many different subject areas. The higher an officer is in the management hierarchy, the less deeply he needs to know about a particular subject, but he will have to be knowledgable in a greater number of subjects.

In the third section, participants are asked about how important they think each of the subject areas is, with regard to enhancing their own knowledge in the subject areas. They are asked whether they think they need to learn more about the subjects. It is anticipated that some (probably those with a positive attitude towards learning) will regard highly the importance of learning as many subject areas as possible. At the same time, some others might feel the need to enhance their knowledge in one or two areas only, or even none at all. By aggregating the scores of all three sections together, it is hoped that a higher order structure of Learning will be measured.

4.3.1 Measurement Scale of the LQ

The precision of response depends on the number of categories of the scales used. The five point Likert scale is probably the most widely used in questionnaires like this, but it might not necessarily be the best for this study. Scales with a higher number of response categories could provide higher precision, simply because it has better potential to discriminate amongst the respondents [27]. In the case of this proposed questionnaire, a five-point Likert scale might be set up such as the following:

1	Very little
2	Little
3	Average
4	High
5	Very high

If a participant chooses Average as his response for his pre-test level of Knowledge (for a particular subject area), he might have difficulty in deciding whether training has successfully helped him increase his level to a High during post-test. An increase from a level to the next might seem too much an achievement. On the other hand, a large number of response categories might lead to difficulties in distinguishing shades of meaning for adjacent responses. As an example is the following bi-polar scale:

> Please indicate your level of knowledge in each subject area by selecting a number from a scale from 1 to 10, where 1 means Very low and 10 means Very high.

With a scale with ten response categories such as this, higher precision can be achieved only if all respondents do not face difficulties in deciding between 1 and 2, 2 and 3, and so on. When they do, there might be inconsistencies in response in repeated measures like pre-and-post tests such as in this study. It is based on these arguments that scales with more than nine categories are not recommended (Fayers *et al.*, pp 34-35). For this questionnaire, it is proposed that a seven-point scale is used for each category. The number of response categories is not too little for an increase to be too much, neither are the categories too narrow for the differences to be too vague.

It can also be anticipated that after attending a course, some participants might give a lower score for his/her level of knowledge, compared to the score given before starting the course. This impression of negative gain in knowledge can be attributed to an increase in awareness of the subject areas themselves. After attending a course, a participant might realise that his knowledge in a particular subject area is actually lower than he had thought. Upon realising that, he or she could give a lower score in the post-test. This increase in awareness could usually be confirmed by the participant's response in section three of the questionnaire - attitude towards the knowledge. After realising that his/her knowledge in a subject area is lower than he/she perceived, the participant would normally indicate a high importance in learning the subject.

A copy of both the pretest and the posttest LQ are included in the appendix section of this thesis. On pages 326 to 330 are the pretest LQ, while the posttest is presented on pages 332 to 336.

4.4 Evaluating the LQ

The LQ is developed based on a very simple model. Ten indicator items are linked to each of the three latent variables or common factors. Wuensch (2005) suggests that confirmatory factor analysis is the right procedure for instances where patterns of relationship between measured variables and common factors is done *a priori*, or before seeing the data [98].

Being newly developed, both the validity and the reliability of the measurement need to be established. As one does not guarantee the other, they need to be assessed separately [67]. Of the three types of validity, *construct* validity is the most relevant as the LQ is intended to measure a psychological attribute. The validity is assessed by confirmatory factor analysis in Section 6.3 on page 101. It would be supported if all items load on their pre-specified constructs.

Each of the Learning factors is measured by ten indicator items, and all ten items should be consistent. This so called internal consistency is one form of reliability and it is most commonly evaluated using Cronbach's coefficient alpha. However, a high degree of internal consistency does not guarantee that a measure is unidimensional ([67], p. 246). The other form of reliability regards aspects of repeatability and stability. Repeatability is the ability of the measurement tool to obtain consistent results under repeated, identical conditions. Stability is the ability of the tool to retain its calibration over a long period of time. Both of these aspects can be evaluated by using different variants of the instruments (equivalent-forms reliability) [27]. For this study, this form of reliability is not tested due to practical constraints.

Chapter 5

Results 1 : Reliability of the Scales

A reliable measurement scale can mean either it is (i) stable over time, or (ii) internally consistent. Of the two, the second one is always desirable [67]. This topic has been discussed in more detailed in Section 3.7.1. In this chapter, we utilize the Cronbach's alpha (Equation 3.3 on page 52) as a measure of internal consistency, and we propose the use of an alternative measure which we call alpha* (Equation 3.8 on page 53). We also calculate the value of ρ (Equation 3.6) for each scale and subscale. Table 5.1 shows the values of Cronbach's coefficient alpha (α), ρ and alpha* of all scales and subscales used in this study.

All of the main scales (GHQ, LQ (pretest and posttest) and CEQ) have alpha values greater than 0.85. The three LQ subscales also show high alpha values. There is not much difference between the alpha values of the pretest and posttest LQ subscales. Among these, the posttest application seems to have the highest reliability (alpha = 0.9126) while the pretest knowledge has the lowest (alpha = 0.8814).

Regarding the GHQ, Goldberg had reported in the *Manual of the GHQ* [35] several reliability coefficients for the GHQ-60. Test-retest reliability coefficients for three different groups of patients were 0.90, 0.75 and 0.51. The value of the split half reliability, where reliability is estimated based on the correlation of two equivalent halves of the scale, was reported to be 0.95.

For the CEQ, even though the alpha for the main scale is 0.8643, not all of the subscales show as high values. Three of them are below 0.8 with the AA showing

Table 5.1: Coefficient alpha, rho and alpha* values of the scales and subscales.

	Coefficient		
Scale/subscales	**alpha**	**Rho**	**Alpha***
General Health (GHQ)	0.8845	0.3896	0.5607
LEARNING (LQ) pretest	0.9343	0.3216	0.4867
Knowledge pretest	0.8814	0.4263	0.5978
Application pretest	0.9054	0.4890	0.6568
Importance pretest	0.9338	0.5852	0.7383
LEARNING (LQ) posttest	0.9512	0.3938	0.5651
Knowledge posttest	0.8853	0.4356	0.6069
Application posttest	0.9126	0.5108	0.6762
Importance posttest	0.9369	0.5976	0.7481
Course Experience (CEQ)	0.8643	0.2169	0.3564
CEQ Good Teaching	0.8730	0.5339	0.6962
CEQ Clear Goals	0.6480	0.3152	0.4793
CEQ Generic Skills	0.8814	0.5533	0.7124
CEQ Appropriate Assessment	0.4866	0.2401	0.3872
CEQ Appropriate Workload	0.7740	0.4613	0.6314

the lowest alpha (0.4866).

Standards of reliability depend on the type of test (Nunnally and Bernstein, pp. 265) [67]. When comparing groups in basic research, an alpha coefficient of 0.80 is adequate. If decisions are to be made about individuals, the minimum level of reliability suggested was 0.90, but Nunnally *et al.* suggested 0.95 as a desirable standard. All the scales in this research are in the first category of test, so the 0.80 adequacy standard is relevant. All but three scales have alphas greater than 0.80. The three which do not are all subscales of the CEQ, namely the CG, the AA, and the AW. The AA which has the lowest alpha (0.4866) has only three items, while both the CG and the AW (alphas 0.6480 and 0.7740 respectively) have four each.

Alpha* values are the reliability values had there been only two items in each scale. As the reliability of a scale increases with the number of items, alpha* values act as standardised reliability measures, making comparisons among the scales possible. Generally all values of alpha* are less than their respective coefficient alphas. None of the scales has alpha* greater than 0.80, with the highest of only 0.7481 (Posttest importance.) There are more consistencies among the LQ subscales, compared to the CEQ subscales which seem to have larger variation in alpha* values.

The main CEQ scale has the lowest alpha* (0.3564) among the main scales. This value is even lower than the lowest among its own subscales, which is the AA with an alpha* of 0.3872.

The fact that the values of alpha* are less than the values of coefficient alpha implies that number of items in the scales makes a significant contribution to the reliability of the scale. It could be suggested that the more the reduction from coefficient alpha to alpha*, the more 'dependent' the scale is on the number of items it consists of. Table 5.2 shows the percentage of reduction of values from coefficient alpha to alpha*.

Table 5.2: Percentage of reduction from coefficient alpha to alpha* values.

Scale/subscales	Alpha value	Alpha*	Reduction(%)
General Health (GHQ)	0.8845	0.5606	36.61
LEARNING pretest	0.9343	0.4867	47.91
Knowledge pretest	0.8814	0.5978	32.18
Application pretest	0.9054	0.6568	27.45
Importance pretest	0.9338	0.7383	20.94
LEARNING posttest	0.9512	0.3938	40.59
Knowledge posttest	0.8853	0.6069	31.45
Aplication posttest	0.9126	0.6762	25.90
Importance posttest	0.9369	0.7481	20.15
Course Experience (CEQ)	0.8643	0.3564	58.76
CEQ Good Teaching	0.8730	0.6962	20.26
CEQ Clear Goals	0.6480	0.4793	26.04
CEQ Generic Skills	0.8814	0.7124	19.17
CEQ Appropriate Assessment	0.4866	0.3872	20.43
CEQ Appropriate Workload	0.7740	0.6313	18.43

The GHQ scale has a reduction of over 36%, while the three LQ subscales, both the pretest and the posttest, have reductions of between 20.15% to 32.18%. The mean reduction of the knowledge subscale is 31.815%, while the figures of the application and importance subscales are 26.675% and 20.545% respectively. This implies that the importance subscale is the least dependent on the number of items.

The CEQ has the largest reduction of 58.76%. This may not be very surprising, as the scale is made up of five different subscales. The subscales are not as bad, having reductions of between 18.43% and 26.04%, with the AW having the lowest reduction. The LQ also consists of different subscales, so reductions from coefficient alphas to alpha* are also quite large for both the pretest and posttest LQ, ie. 47.91% for the pretest and 40.59% for the posttest.

5.0.1 Discussion on coefficient alpha (α), rho (ρ), and alpha* (α^*)

As mentioned previously in this chapter and in Section 3.7.1, Cronbach's coefficient alpha (α) indicates internal reliability of a measure, which relates to a latent variable **L**. The latent variable is measured adequately if α is high enough. However, α may be high either because there are many items in the scale weakly related to **L**, or because there are few items in the scale highly related to **L**. Where possible, the latter is always preferred.

The simplest assumption is that each item in the scale has the same correlation, rho (ρ), with **L**. We prefer ρ to be high. Equation 3.5 on page 52 shows that ρ and α are related via the number of items, k, in the scale. If α_1 and α_2 are alpha values for two scales and both scales have equal underlying correlation ρ, then $\alpha_1 < \alpha_2$ if $k_1 < k_2$. From another perspective, the more items in the scale, the lower the implied correlation for a fixed value of α. Many sociological analyses ignore this fact, and report only α. However, we feel that it is important also to report the implied underlying correlation.

As an alternative way to thinking about ρ, alpha* (α^*) is provided for people used to the sociological literature, and used to seeing reliabilities reported, rather than implied correlations. Its advantage is that it is a reliability figure, but also standardized, in the sense that α_1^* and α_2^* from two studies may be directly compared without referring to the number of items in the scale. A low α^* implies that the scale depends on many items in order to achieve acceptable reliability. On the other hand a high α^* implies that few items are needed for acceptable reliability.

5.1 Individual Analysis of Items in the Scales

In each scale, each of the items that make up the scale is analysed for its contribution to the reliability of the scale. Taking the item out of the scale, Cronbach's alpha is calculated with the remaining $k-1$ items. This is then repeated with the next item and so on. The results are presented in Table 5.3 through Table 5.5.

Table 5.3: Values of alpha if the item is omitted from the GHQ scale (α = 0.8845)

Item	Alpha	Item	Alpha
Item 1	0.8797	Item 7	0.8759
Item 2	0.8776	Item 8	0.8783
Item 3	0.8797	Item 9	0.8674
Item 4	0.8776	Item 10	0.8712
Item 5	0.8719	Item 11	0.8770
Item 6	0.8759	Item 12	0.8688

Table 5.4: Values of alpha if the item is omitted from the pretest LQ subscales.

	Knowledge	Application	Importance
Item	(α=0.8814)	(α=0.9054)	(α=0.9338)
Item 1	0.8682	0.8918	0.9246
Item 2	0.8719	0.8959	0.9266
Item 3	0.8845	0.8997	0.9297
Item 4	0.8745	0.8989	0.9272
Item 5	0.8622	0.8894	0.9228
Item 6	0.8688	0.8942	0.9273
Item 7	0.8663	0.8964	0.9271
Item 8	0.8668	0.8960	0.9263
Item 9	0.8667	0.8946	0.9285
Item 10	0.8665	0.8986	0.9299

Results for the GHQ in Table 5.3 suggest that omitting any item from the scale does not increase the coefficient alpha. Neither does it significantly decrease the alpha. The smallest coefficient shown as a result of omitting an item is related to item 9.

Table 5.4 shows that for the knowledge subscale, omitting item 3 increases the overall coefficient alpha by a mere 0.0031. No other item of the subscale shows a similar effect if taken out. For the application subscale, there is not much difference in the value of coefficient alpha when any item is omitted, though all indicate reductions. The observation is similar with the importance subscale.

Results for the CEQ in Table 5.5 are more varied. A general observation is that no omission of an item causes the coefficient alpha of the respective subscale to

Table 5.5: Values of alpha if the item is omitted from the CEQ subscales.

	Good	**Clear**	**Generic**	**Appropriate**	**Appropriate**
	teaching	**goals**	**skills**	**assessment**	**workload**
Item	(α=0.8730)	(α=0.6480)	(α=0.8814)	(α=0.4866)	(α=0.7740)
Item 1	0.8537	0.5852	0.8565	0.4079	0.7316
Item 2	0.8606	0.5047	0.8585	0.3178	0.7946
Item 3	0.8510	0.6393	0.8665	0.4289	0.6384
Item 4	0.8590	0.5896	0.8434		0.6927
Item 5	0.8402		0.8797		
Item 6	0.8417		0.8601		

increase, except for item 2 of the AW subscale. However, the AW subscale has only 4 items, thus taking one out decreases the domain and might affect the validity. Subscales with less number of items seem to be more affected by omission of an item. This is obvious with the CG, AA, and AW subscales, while for the GT and the GS subscales, each with six items, the reduction in the coefficient alpha is not as much.

5.2 Normality of the CEQ Scales

The CEQ consists of five subscales, each of which is indicated either by 3, 4, or 6 indicator items. All of the indicators are measured using a common five point scale. Since the number of indicators for the five subscales are not the same, the means are taken as the scores rather than the sums. Thus the scores range from 1 to 5.

Normality of these subscales are assessed by looking at their normal probability plots as well as their histograms. A 'fat pencil test' is then applied to the plots to see whether they indicate deviation from a Normal distribution. Normal probability plots of all CEQ scales will display more-or-less granularity because of their construction from Likert scales.

The GT subscale of the CEQ consists of six items. Figure 5.1 shows that the distribution is slightly left-skewed with some outliers in the tail. Granularity is also obvious.

The CG subscale consists of four items. Figure 5.2 indicates long tails at both ends. Otherwise the distribution is reasonably normal.

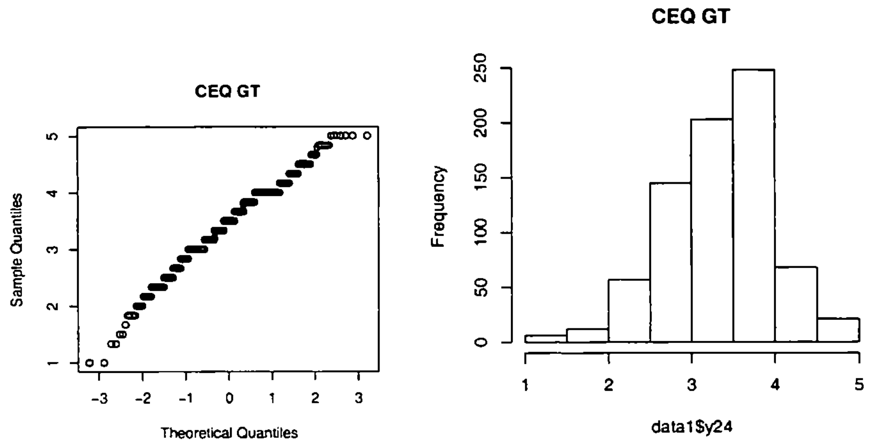


Figure 5.1: CEQ Good Teaching

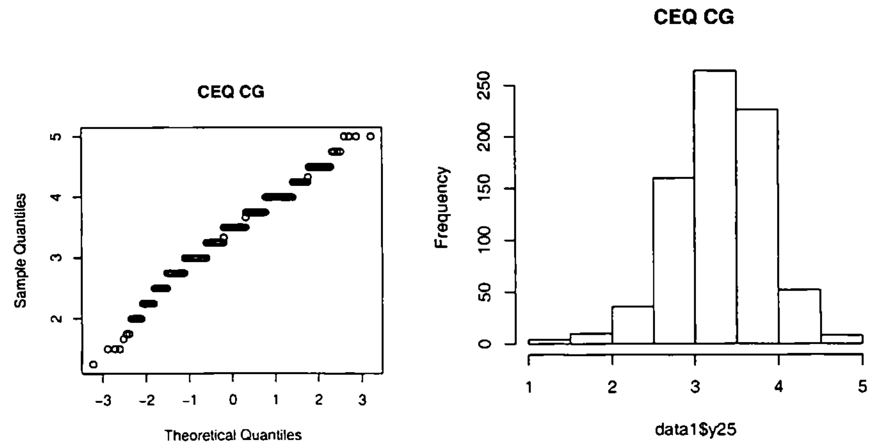


Figure 5.2: CEQ Clear Goals

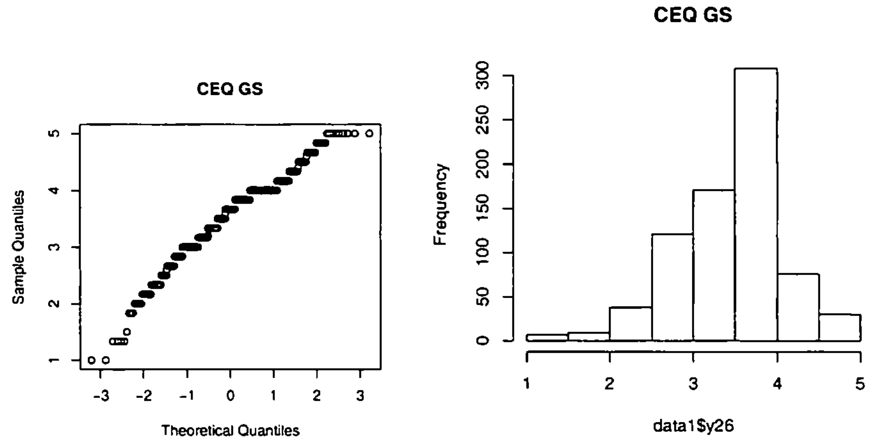


Figure 5.3: CEQ Generic Skills

The GS is a scale with six items. It has a long tail at the lower end of the scale, but otherwise the distribution is roughly Normal, except for the ceiling effect. The AA subscale has only three indicator items. The distribution in Figure 5.4 looks reasonably Normal, with fat tails and but with even more granularity.

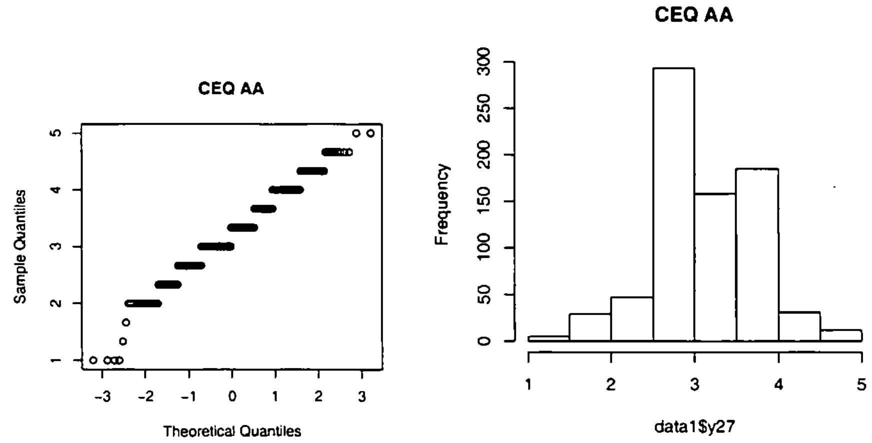


Figure 5.4: CEQ Appropriate Assessment

The AW subscale consists of four items. It does not seem to have significant outliers, and Figure 5.5 shows that the distribution is roughly normal, but with fat tails again. Overall, all subscales of the CEQ seem to be approximately Normally

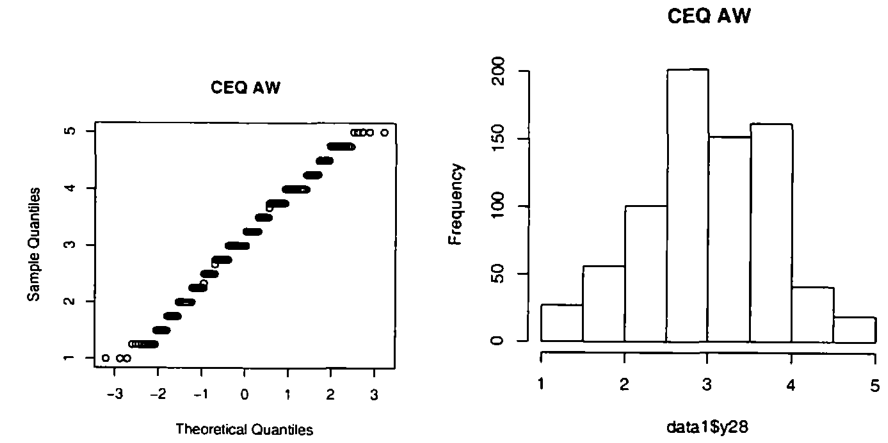


Figure 5.5: CEQ Appropriate Workload

distributed. There are however indications for outliers, and for fatter tails than normal.

5.3 Normality of the GHQ Scale

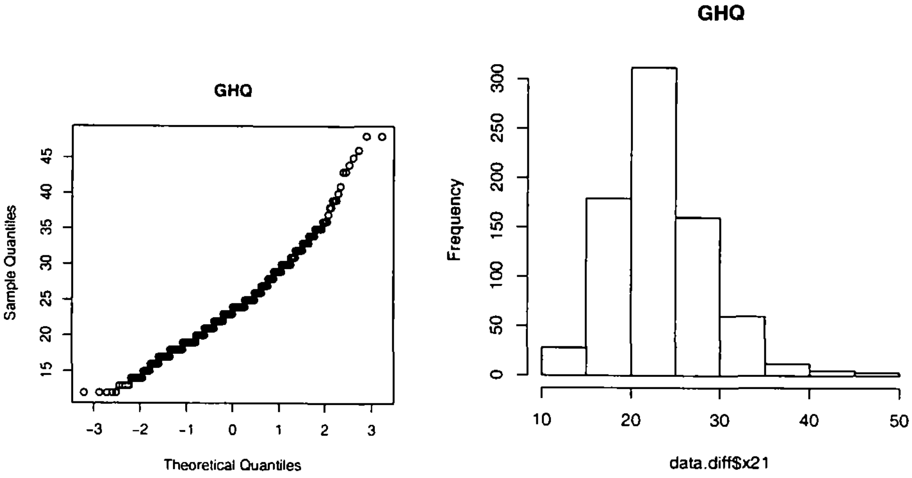


Figure 5.6: General Health

The GHQ is a scale consisting of twelve items, each evaluated on four point scale.

The score thus is in the range of 12 through 48. Figure 5.6 shows a slight deviation from Normality, but not excessive. The histogram also suggests approximation to Normal distribution, except for some values at the upper tail.

5.4 Normality of the LQ Subscales

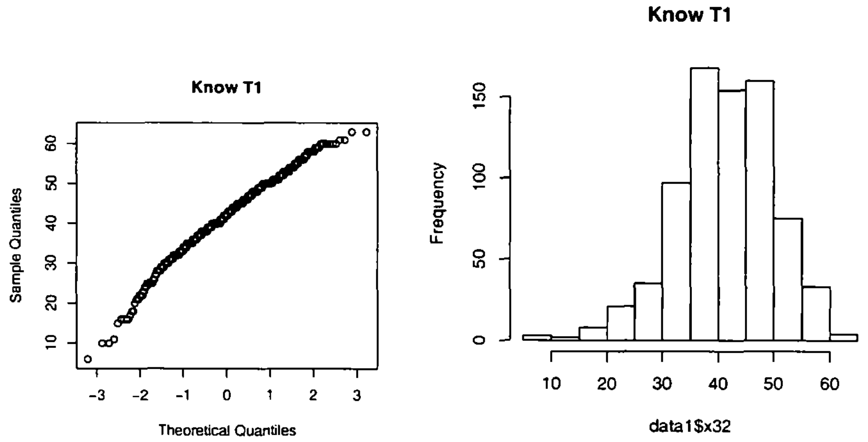


Figure 5.7: Pretest Knowledge

Figure 5.7 indicates that pretest knowledge seems to be slightly left-skewed. Effects of outliers can also be seen at the lower end of the scale. Otherwise the distribution is roughly Normal. At time 2, the tail at the lower end is longer, as shown in Figure 5.8. The distribution is now more skewed, with an increase in the number of outliers at both ends. The middle part of the distribution however stays roughly Normal.

Application time 1 (Figure 5.9) does not seem to be Normally distributed. There is a clear ceiling effect, as a result of many respondents giving top scores. The same conclusion is also applicable to application time 2, as presented in Figure 5.10.

Figure 5.11 shows that pretest importance is skewed to the left. It has a long tail at the lower end, and a large number of observations at the maximum score. The situation for posttest importance is not much different from its pretest scores, as can be seen in Figure 5.12. If anything, the left tail is longer, probably because of an increase in the number of outliers.

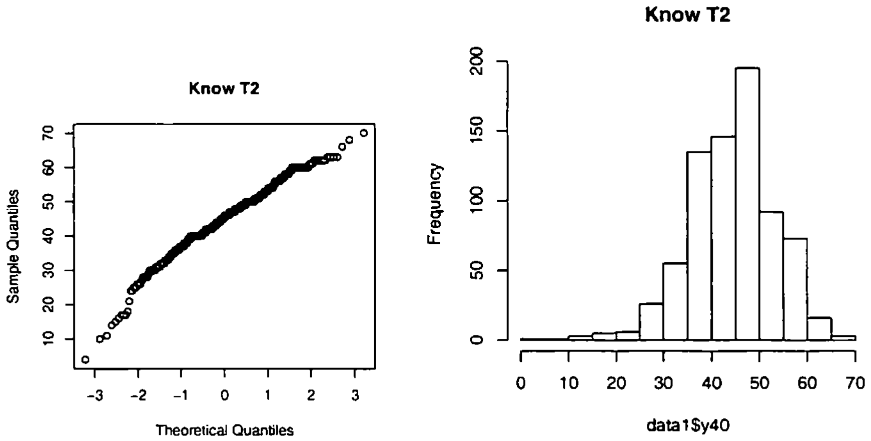


Figure 5.8: Posttest Knowledge

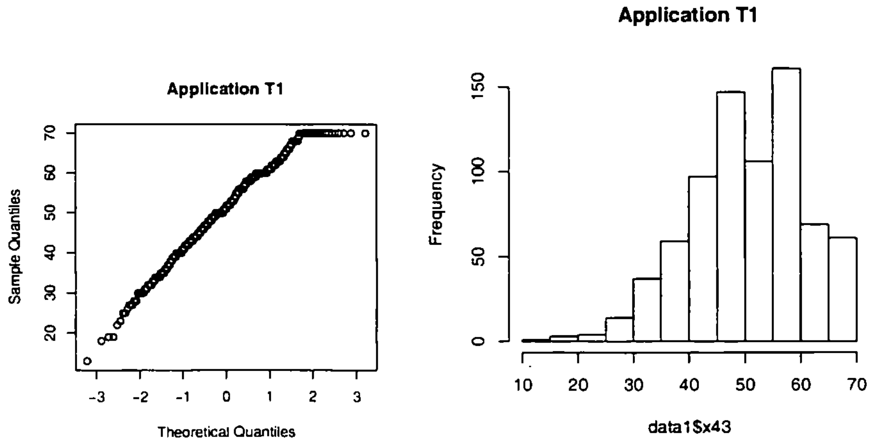


Figure 5.9: Pretest Application

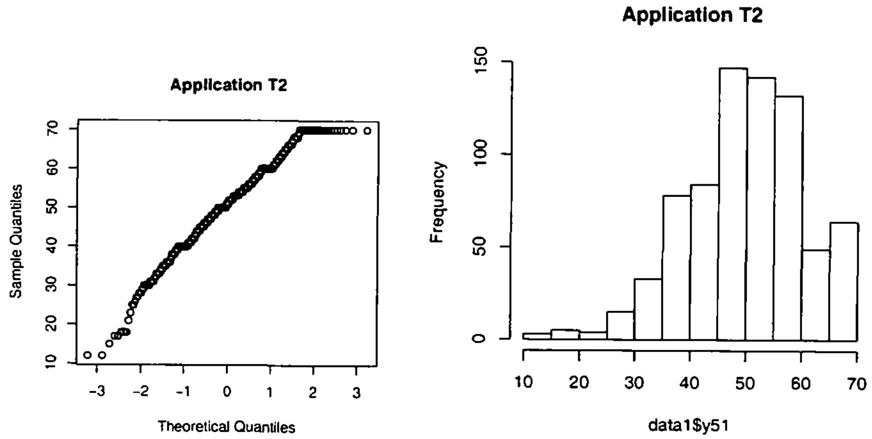


Figure 5.10: Posttest Application

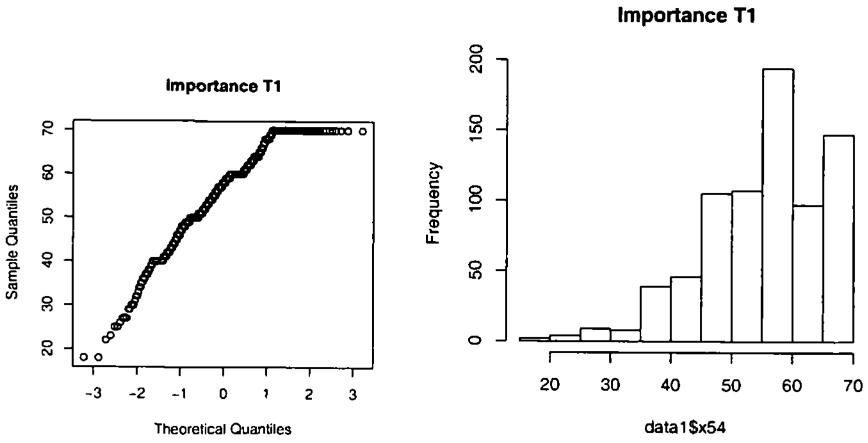


Figure 5.11: Pretest Importance

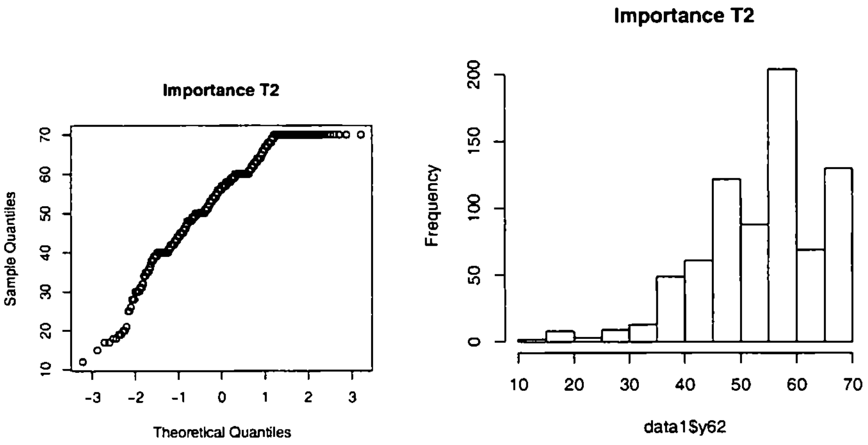


Figure 5.12: Posttest Importance

5.5 Conclusions of Reliability Analyses

Most of the scales used in this study have high values of Cronbach's coefficient alpha. Only three subscales, all of the CEQ, have alpha less than 0.80. Analysis of reliability when one item is omitted shows that in most scales, there is no single item which is detrimental to the scales' reliability. The only exceptions are item 3 of the pretest knowledge and item 2 of the appropriate workload, where omission of the scale results in a slight increase in the coefficient alpha.

The CEQ scale also has the largest percentage of reduction from coefficient alpha to alpha*, indicating a relatively high 'dependency' on having a large number of items. This large reduction may be attributed to the scale consisting of five different subscales, caused by high variation and low reliability when all the indicator items are grouped together. The LQ scale, which has three subscales, also has a large reduction from coefficient alpha to alpha* for both the pretest and the posttest scores. Reductions of their respective subscales are smaller than the main scales. This is easily understood, as the indicator items within a subscale are consistent and highly related. The GHQ scale, which is a unidimensional measure, has the lowest reduction percentage.

Generally, all subscales of the CEQ are approximately Normally distributed. The GHQ does not deviate too much from Normal distribution as well. Of the three subscales of Learning, only knowledge shows rough approximation to Normality. The other two subscales, the application and the importance, indicate deviations from a Normal distribution, mainly because of the ceiling effect in the scores.

Chapter 6

Results 2 : Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) on the models of the measurement tools of this study is done using Structural Equation Modeling (SEM) [12]. The main objective is to test whether the model of each of the measurement scales conforms to the data. All three measurement scales, the Course Experience Questionnaire (CEQ), the General Health Questionnaire (GHQ) and the Learning Questionnaire (LQ) are tested separately.

The models of the CEQ and the LQ have two levels: the level of the measurement models and the level of the structural model. The CEQ has five measurement models, while the LQ has three. Measurement models are examined first before tests on structural models make any sense. The GHQ has only a measurement model and no structural model. The SEM tests are carried out using the AMOS 6 software.

6.1 Evaluation of the CEQ

The five subscales of the CEQ are Good Teaching (GT), Clear Goals (CG), Generic Skills (GS), Appropriate Assessment (AA), and Appropriate Workload (AW). Each of these models is tested for model fit before the structural model of the CEQ is examined. The complete CEQ model is as presented in Figure 6.1.

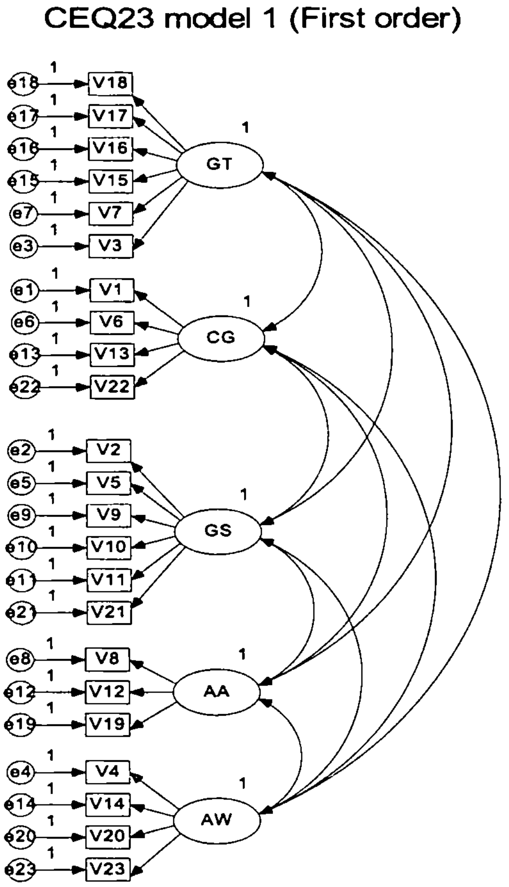


Figure 6.1: First order CEQ model

Just like in any other SEM models in this thesis, the indicators in the CEQ model are assumed to be continuous and having two causes. One cause is the underlying subscale that they are supposed to measure, and the second cause is a combination of all other sources represented by the error term [54]. The errors are also assumed to be independent of each other and of the underlying subscale.

All measurement errors in the model are assigned a scale through a unit loading identification (ULI) constraint. This gives the unstandardised residual path coefficient of a measurement error on the indicator a value of 1.0. Consequently the measurement error has a scale related to that of the unexplained (unique) variance of its indicator.

AMOS reports measures of fit for 3 types of models: the default model (user specified model), the saturated model, and the independence model. *The saturated* *model* is a perfect model (Ullman, 2001 [90]). This is the most general model possible, where no constraints are placed on the population moments. In one sense, it is guaranteed to fit any set of data perfectly. The *independence model* on the other hand is the other extreme. Observed variables are assumed to be uncorrelated with each other. The user defined model is taken to be lying somewhere between these two extreme models [5]. For this study, values for default model is reported.

The results of the SEM analyses on all CEQ models are presented in Table 6.1. The discussions of the results are in the subsections that follow.

Table 6.1: Results of Course Experience measurement models assessment.

	Subscales					**Good fit**
Indices	*GT*	*CG*	*GS*	*AA*	*AW*	
Model χ^2	100.489	0.557	44.012	NA	8.225	
Df	9	2	9	1	2	
P-value	0.00	0.757	0.00	NA	0.016	Non-sig
CMIN/DF	11.165	0.279	4.89	32.895	4.128	< 2
RMR	0.026	0.004	0.015	0.053	0.020	Close to 0
GFI	.956	1.000	0.980	0.971	0.994	Close to 1
AGFI	.897	0.998	0.954	0.827	0.972	Close to 1
PGFI	.410	0.200	0.420	0.162	0.199	Close to 1
NFI	.950	0.999	0.979	0.741	0.991	$> .9$
CFI	.954	1.00	0.983	0.743	0.993	$> .95$
RMSEA	.118	0.000	0.073	0.208	0.065	< 0.05

6.1.1 Good Teaching (GT)

The Good Teaching (GT) subscale consists of 6 indicator items. The items are questions 3, 7, and 15 through 18 of the CEQ scale. To scale the GT factor, unit variance identification (UVI) constraint is imposed by fixing the factor variance to 1.0 and effectively standardising the factor. UVI is more common than the ULI (Kline, pp. 171) and as an effect all factor loadings for the factor's indicators are free parameters.

Identification. A model is identified when (1) the number of free parameters is less than or equal to the number of observations, and (2) every latent variable has a scale [54]. In this model, the number of observations is $v(v+1) = 6(6+1) = 42$. Parameters are made up of 6 variances (of the 6 measurement errors) and 6 direct effects (of factor on indicators), giving a total of 12. Degrees of freedom for the model is the difference between the distinct sample moments[1], or the number of data points [90], and the number of parameters. In this case, the distinct sample moments is $(6(6 + 1))/2 = 21$, and the number of parameters is 12, giving 9 as the degrees of freedom.

Results. Indices such as RMR, GFI, NFI, and CFI show values that indicate good fit. At the same time, other indices such as CMIN/DF, AGFI, PGFI, and RMSEA show values that do not indicate this measurement model as fitting the data well. Overall, it can be concluded that the measurement model of the GT roughly fits the data.

6.1.2 Clear Goals (CG)

The Clear Goals (CG) subscale is indicated by 4 observed variables. The indicators are items 1, 6, 13 and 22. In this model, the number of observations is $v(v + 1)= 4(+ 1)= 20$. There are 4 variances of the measurement errors and 4 direct effects of the Clear Goals subscale on the indicators, making a total of 8 parameters. Degrees of freedom for the model equals $(v(v + 1))/2$ minus the number of parameters, which is 10 - 8 = 2.

Results. Most indices show values of well-fitting model. The only indices which do not are CMIN/DF and PGFI. It is thus concluded that the measurement model of the CG fits the data well.

6.1.3 Generic Skills (GS)

The GS subscale consists of 6 items, namely items 2, 5, 9 to 11 and 21. The identification for the model is similar to that of the GT model, with 9 degrees of freedom.

Results. Indices that show values of well-fitting model for the GS are RMR, GFI, AGFI, NFI, and CFI. It is concluded that the measurement model of GS fits the data well.

[1] Terms used by AMOS

6.1.4 Appropriate Assessment (AA)

The Appropriate Assessment (AA) subscale has only 3 indicators, the least among the CEQ subscales. In this model there are 3 observed variables (the indicators) and 6 parameters to be estimated (the 3 loadings and the variances of the measurement errors). The number of observations is thus v(v + 1)= 3(3 + 1)=12. Degrees of freedom is 12/2 - 6 = 0.

Because the degrees of freedom is zero, many of the model fit statistics either could not be computed or does not give proper readings. Therefore, one of the direct effects is given a fixed value of 1.0. The effect is that the degrees of freedom is not zero but 1, making calculations for many of the fit statistics possible.

Results. Two indices show values of a well-fitting model, namely RMR and GFI. The index of AGFI shows a slightly less that well-fitting value, as do indices of NFI, CFI and RMSEA. The conclusion for this measurement model is that it does fit the data, but the fitting is not very good.

6.1.5 Appropriate Workload (AW)

The Appropriate Workload (AW) subscale has 4 indicators, which are items 4, 14, 20 and 23. Identification is similar to the Clear Goals, with 2 degrees of freedom.

Results. Almost all indices show values of a well-fitting model. The only indices which show values of a slighty less well-fitting model are CMIN/DF, PGFI, and RMSEA. It is concluded that the measurement model of the AW fits the data well.

The overall observation for the CEQ subscales is that different indices suggest different conclusions. For each of the measurement models, there are indices that show values of a well-fitting models and there are indices that show lack of fit. However in most cases, there are more indices that show the models do fit the data than indices that do not.

The index of PGFI and the ratio of CMIN over DF almost never show good fit. On the other hand, indices like RMR, GFI, NFI and CFI do show good fit in most cases. CMIN/DF is based on the chi square value, which tends to be easily significant when the sample is large, as in this case. It is therefore no surprise that the index always shows lack of fit in terms of the models.

The seemingly worst performing measurement model is the AA. It's model χ^2 cannot be calculated and it also has a very high score of CMIN/DF index. Other indices also indicate values far from good fit indications. This could due to the fact that this measurement model only has 3 indicators, the least among the CEQ subscales. On the other hand, the measurement model of the CG seems to be the best performing, with indices showing values close to good-fit indications.

6.1.6 CEQ Structural Model

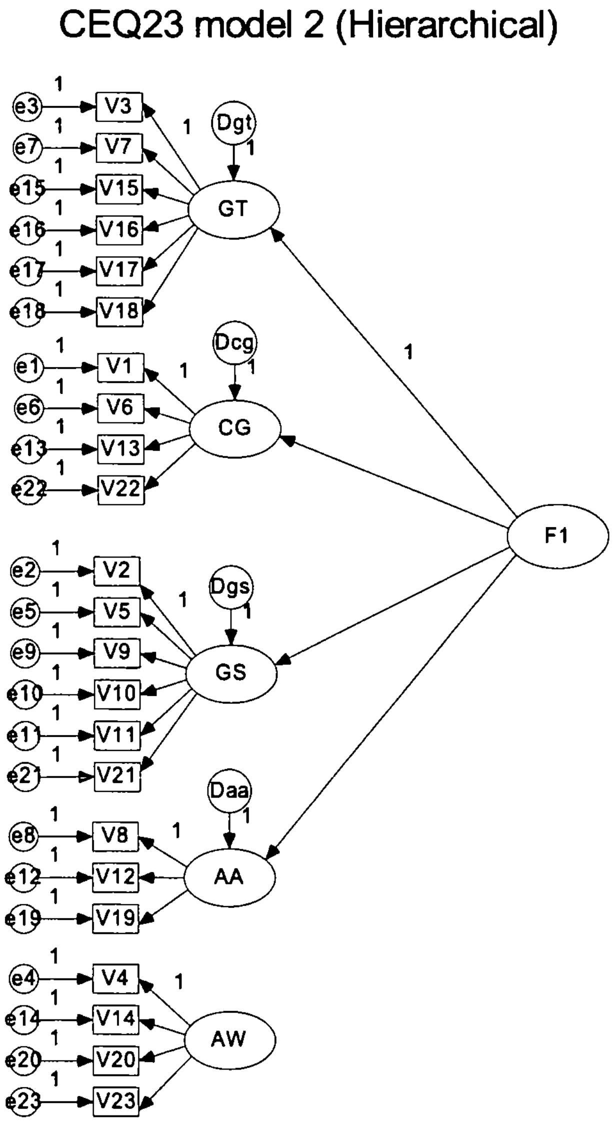
Two structural models of the CEQ are tested. The first is as in Figure 6.1 on page 90. This is the standard confirmatory factor analysis model, where all five factors are suggested to covary with each other [54]. For this analysis, identification is achieved by fixing the variance of each factor to unity, and letting all factor loadings to be free variables.

In the model there are 23 observed variables giving $(23(23 + 1))/2 = 276$ observations (data points). As there are 56 parameters to be estimated (23 variances, 23 direct effects and 10 covariances), the test is done with 276 - 56 = 220 degrees of freedom. The model fit tests results are in the table below.

Course Experience model 1 results.

Indices	**Value**	**Good fit.**
Model χ^2	1122.688	
P-value	0.00	Non-sig
CMIN/DF	5.103	< 2
RMR	.065	Close to 0
GFI	.881	Close to 1
AGFI	.851	Close to 1
PGFI	.703	Close to 1
NFI	.854	$> .9$
CFI	.878	$> .95$
RMSEA	.075	< 0.05

Results show that none of the fit statistics shows values of a well-fitting model. Indices such as RMR, GFI, AGFI and NFI show values slighly less than the thresholds



CEQ23 model 2 (Hierarchical)

Figure 6.2: CEQ23 hierarchical model.

of good-fit. All these indicate that the model does not fit the data very well.

The second model tested is the CEQ model with suggested second order factors (Please refer to Figure 6.2 on page 95). This model is suggested by Wilson *et al* in 1997 [96]. The first 4 factors (GT, CG, GS, and AA) are hypothesised to indicate one higher factor, while the AW indicates another factor. The four factors have a common direct cause and this implies that they do not have direct associations among themselves but exists only through the factor.

Standardizing the second-order factor by fixing its variance to 1.0 is one option of scaling it, but it is not recommended for multiple sample analysis [54]. In this case, one of the direct effects is given a fixed value of 1.0 to assign scale to the factor.

The five first-order factors are now endogenous, thus their variances can no longer be fixed. Each of them has a disturbance as a unique variable, and each of these disturbances is given a fixed value of 1. To identify the model, one direct effect from the factor to one of the first-order factors is fixed to 1.

In the original suggestion ([96]), the AW subscale indicates another higher-order factor by itself. In the model, it is not possible to have a higher-order factor with a direct effect on the AW because that would make the whole model unidentified. In this analysis, the AW factor is just assumed to be uncorrelated with the second-order factor.

In this model, there are $(23(23 + 1))/2 = 276$ data points. The parameters to be estimated include 29 variances (of 23 measurement errors, 5 disturbances and 1 second-order factor) and 21 direct effects (18 on indicators and 3 on first-order factors). Thus the degrees of freedom for this model is 276 - 50 = 226. Model fit is as in the following table.

Course Experience model 2 results.

Indices	**Value**	**Good fit.**
Model χ^2	1210.845	
P-value	0.000	Non-sig
CMIN/DF	5.358	< 2
RMR	.075	Close to 0
GFI	.869	Close to 1
AGFI	.840	Close to 1
PGFI	.712	Close to 1
NFI	.842	$> .9$
CFI	.867	$> .95$
RMSEA	.077	< 0.05

This model is no better than the first one in term of fitting the data. The fit indices show values which are not very different from the values for the first model. Similar to the first model, this hierarchical model does not fit the data well, but the fit is not very bad either. This lack of fit of the two CEQ structural models may not be attributed to the large sample size. This is because the indices that show lack of fit include not only χ^2 and CMIN/DF, but also other indices which are not related to the chi-square statistics.

6.2 Evaluation of the GHQ

The GHQ is a straightforward one factor first order model with twelve indicators. The model is presented in Figure 6.3.

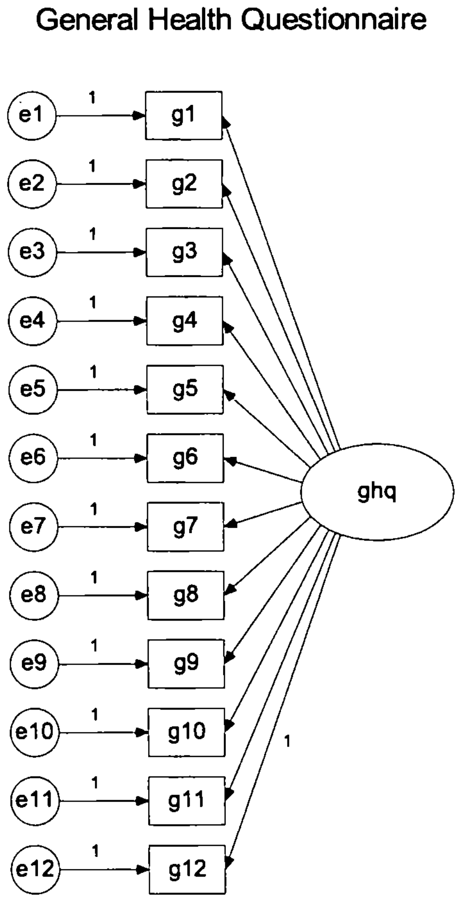


Figure 6.3: General Health model

For identification purposes, a direct effect from the general health factor to indicator number twelve is fixed to 1. This gives the factor the same scale as the indicator. There are twelve observed variables, thus this model has (12(12 + 1))/2 = 78 distinct sample moments or data points. The parameters to be estimated are twelve error terms of the indicators, eleven direct effects from the general health factor to the indicators, and the variance of the factor, making a total of twenty-four. The model thus is tested at 78 - 24 = 54 degrees of freedom.

6.2.1 SEM on the GHQ

Part of the output from the SEM analysis on the GHQ data is presented in Table 6.2. Because the results of the analysis are not very good, four of the largest **modification indices** (MI) are included in the table.

As explained earlier in section 3.8.5 on page 60, each index of the MI reflects the predicted decrease in χ^2 value if the parameter or equality constraint is removed from the model and the model is re-estimated. Values under the 'Par change' in Table 6.2 column are estimates of how much the coefficient would change.

For MI that relates to the covariances, it has to do with the decrease in χ^2 if the two error terms are allowed to correlate. In the case of regression weights, the MI has to do with the decrease in χ^2 if the path between the two variables is added on.

Table 6.2: Model fit and modification indices of General Health.

	Model fit	—	*Modification indices.*		
Indices	**Value**	**Good fit.**	**Covariances**	**M.I.**	**Par change**
Model χ^2	1092.256		e4 ↔ e3	160.376	.101
P-value	0.000	Non-sig	e8 ↔ e4	130.909	.086
CMIN/DF	20.227	< 2	e9 ↔ e5	121.957	.131
RMR	.040	Close to 0	e11↔e10	105.642	.094
GFI	.756	Close to 1	**Regression weights**		
AGFI	.647	Close to 1	g4 ← g3	108.540	.327
PGFI	.523	Close to 1	g3 ← g4	102.666	.313
NFI	.732	$> .9$	g4 ← g8	85.912	.305
CFI	.741	$> .95$	g8 ← g4	83.822	.267
RMSEA	.161	< 0.05			

The first part of the MI suggests adding covariances between error terms 3 and 4, 4 and 8, 5 and 9, and 10 and 11. These suggested covariances violate the assumption that the error terms are independent of each other. The second part of the MI suggests adding paths between two pairs of indicators, namely indicators 3 and 4, and indicators 4 and 8. These suggestions imply high correlation between the indicators. They do not violate any assumptions, but modifications need theoretical justifications. Values of their correlations coefficients are presented in Table 6.3.

Table 6.3: Correlations coefficients of GHQ indicators.

Indicators	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00											
2	0.42	1.00										
3	0.34	0.26	1.00									
4	0.43	0.18	**0.61**	1.00								
5	0.41	0.50	0.23	0.26	1.00							
6	0.15	0.29	0.22	0.27	0.42	1.00						
7	0.45	0.36	0.26	0.31	0.45	0.24	1.00					
8	0.40	0.17	0.44	**0.60**	0.16	0.32	0.30	1.00				
9	0.41	0.48	0.28	0.32	0.66	0.44	0.49	0.28	1.00			
10	0.29	0.33	0.47	0.41	0.39	0.55	0.25	0.41	0.50	1.00		
11	0.21	0.29	0.50	0.41	0.33	0.42	0.22	0.38	0.37	0.59	1.00	
12	0.54	0.40	0.40	0.46	0.44	0.30	0.54	0.46	0.63	0.39	0.35	1.00

Correlations between indicators 3 and 4 (0.61) and indicators 4 and 8 (0.60) are indeed among the highest, but their values are not the largest. Indicator 5 is correlated with indicator 9 with a value of 0.66, but this pair is not observed among the largest four MI.

6.2.2 Conclusions of Analysis on the GHQ.

Based on the fit statistics, the GHQ model does not fit the data well. Chi square based statistics (χ^2 and CMIN/DF) are showing values nowhere near good fit.

The values of other statistics are not good either. The Root Mean Square (RMR), which calculates the average difference between the sample variance and covariance matrix and the estimated population's equivalence, is probably the only one that shows a value not too far off from a good fit. The other indices however are showing values which are far from the thresholds of a well-fitting model.

The four largest modification indices are shown in Table 6.2. For direct effects, MI suggest that item 4 is dependent on item 3 and vice versa. Similarly item 4 is suggested to be dependent on item 8 and vice versa. Referring back to the questionnaire, item 4 is a question specifically about '*making decisions about things.*'. Item 3 meanwhile is about '*playing a useful part in things.*' It could have been that these two items looked too similar to each other to be effectively differentiated by the respondents.

A scree plot of the GHQ (Figure 6.4) shows a dominant single factor. Three

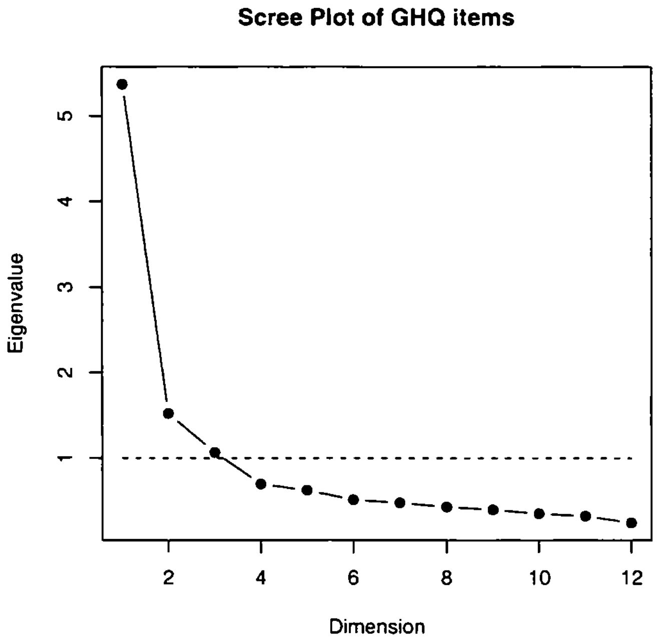


Figure 6.4: Scree plot of the General Health data

factors have eigenvalues greater than 1.0, but only one of them seems to be the main factor. The other two are most probably just random and negligible effects. A check on the scale's factor analysis with three factors gives loadings as in Table 6.4.

Table 6.4: Factor loadings of General Health data with 3 factors

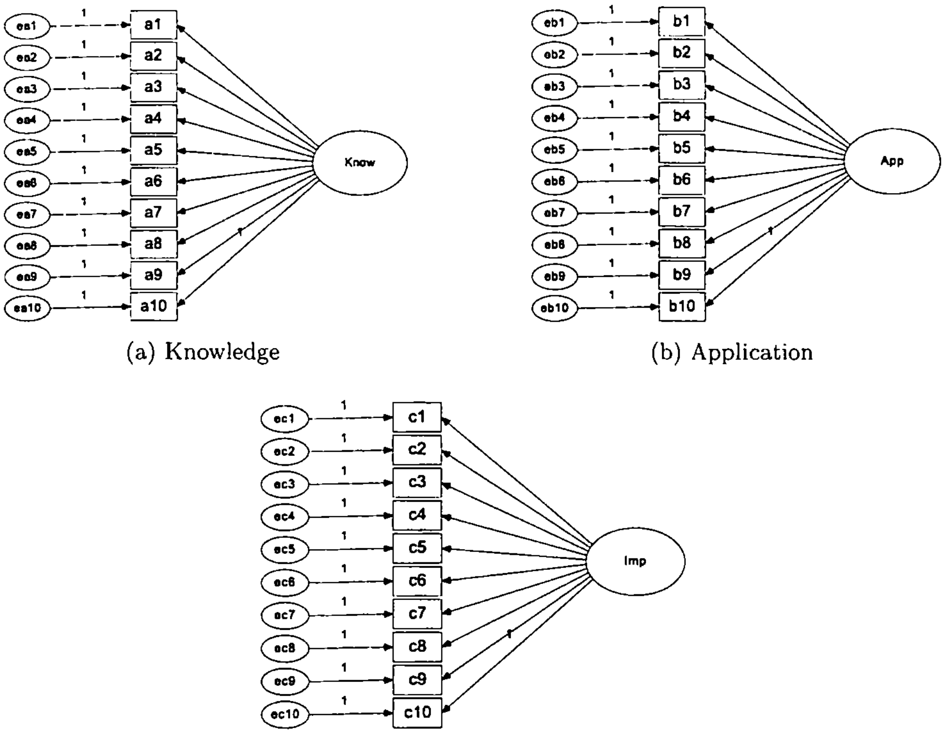
Item no.	Factorl	Factor2	Factor3
1	**0.471**	0.347	-
2	**0.552**	-	0.304
3	0.126	**0.650**	0.313
4	0.190	**0.784**	0.186
5	**0.725**	-	0.311
6	0.316	0.187	**0.589**
7	**0.604**	0.277	-
8	0.204	**0.642**	0.233
9	**0.779**	-	0.369
10	0.290	0.317	**0.713**
11	0.156	0.386	**0.611**
12	**0.699**	0.366	0.130

The factor loadings show that not all twelve items of the scale load on one factor. Items 3, 4 and 8 had highest loadings on factor 2, while items 6, 10 and 11 also load on factor 3. However, not much should be read from the loadings on factor 2 and factor 3. Each item that loads on those factors also loads on factor 1, albeit with smaller coeffficient. Overall, it can be concluded that the GHQ model has one dominant factor, and two random factors. This finding is displayed graphically by the scree plot previously.

6.3 Evaluation of the LQ

The LQ scale is made up of three subscales: knowledge application and importance. The measurement was done twice, before training (time 1) and after training (time 2). Each of the subscales is indicated by ten items, referring to the ten subject areas related to the training. Before the structural model of Learning can be examined, each of the measurement models has to be tested for goodness-of-fit. These models are presented in Figure 6.5.

In each of these models, there are ten observed variables and one factor. The



(a) Knowledge

(b) Application

(c) Importance

Figure 6.5: Knowledge, Application and Importance measurement models.

number of data points or distinct sample moments for each model is (10(11))/2 = 55. One of the direct effects is fixed to 1.0 to identify the model. The parameters to be estimated are nine direct effects and eleven variances (of ten error terms and one factor), making a total of 20. Tests are therefore done at 35 degrees of freedom.

To evaluate the measurement models, the structural equation modeling (SEM) is used. This analysis determines whether each of the model fits the data. If the fit is good, then we can proceed to examining the LQ structural model. Otherwise, if the fit is not good, then the modification indices (MI) will suggest modification to the models.

6.3.1 Results of SEM on the LQ Subscales

The results of the SEM analyses on the measurement models of the LQ are presented in Table 6.5. It is obvious that all three measurement models do not fit the data at

Table 6.5: Summary of LQ measurement models assessment.

	Pretest.			Posttest.			
Indices	*Know*	*App*	*Imp*	*Know*	*App*	*Imp*	**Good fit.**
Model χ^2	1830.84	2069.81	2341.823	1893	2233.67	2584.63	
Df	35	35	35	35	35	35	
P-value	0.00	0.00	0.00	0.00	0.00	0.00	Non-sig
CMIN/DF	52.31	59.137	66.909	54.108	63.819	73.846	< 2
RMR	0.154	0.189	0.141	0.159	0.165	0.149	Close to 0
GFI	0.697	0.695	0.654	0.647	0.650	0.591	Close to 1
AGFI	0.524	0.521	0.456	0.446	0.450	0.357	Close to 1
PGFI	0.444	0.443	0.416	0.412	0.414	0.376	Close to 1
NFI	0.562	0.600	0.656	0.569	0.607	0.645	$> .9$
CFI	0.566	0.604	0.659	0.573	0.610	0.648	$> .95$
RMSEA	0.276	0.293	0.312	0.280	0.305	0.328	< 0.05

all, at both time points. All indices show values very far off from the thresholds of a good-fit. All measurement models of the LQ do not fit the data at all. The findings are the same for both pretest and posttest data.

Since the results show bad fit of the models, we now examine the MI. Covariances and regression weights with MI greater than 100 ([34]) are presented. The MI suggest modifications to the models by adding covariances, or adding direct paths between indicators. For each modification, the estimated reduction in the χ^2 value is given as the MI index. The estimated value of the covariance, or of the direct path is given as the Parameter Change (Par. chg.) in the MI output tables. However, these modifications should only be done only if there is statistical or theoretical sense to them.

Table 6.6: MI of Knowledge measurement model.

Pretest			*Posttest*		
Covariances	**M.I.**	**Par chg.**	**Covariances**	**M.I.**	**Par chg.**
ea2 ↔ ea1	222.076	.509	ea2 ↔ ea1	292.953	.430
ea6 ↔ ea5	127.905	.409	ea6 ↔ ea5	156.259	.431
ea8 ↔ ea7	196.571	.560	ea8 ↔ ea7	251.960	.636
ea10↔ ea9	648.520	.704	ea10 ↔ ea9	515.511	.521
Reg. weights			**Reg. weights**		
a1 ← a2	141.512	.400	a1 ← a2	146.276	.369
a2 ← a1	124.234	.333	a2 ← a1	141.927	.343
a9 ← a10	280.799	.468	a7 ← a8	118.676	.322
a10← a9	281.210	.488	a8 ← a7	116.610	.309
			a9 ← a10	278.832	.507
			a10 ← a9	284.363	.523

Table 6.7: MI of Application measurement model.

Pretest			*Posttest*		
Covariances	**M.I.**	**Par chg.**	**Covariances**	**M.I.**	**Par chg.**
eb2 ↔ ebl	188.938	.420	eb2 ↔ ebl	240.059	.372
eb6 ↔ eb5	138.707	.463	eb6 ↔ eb5	213.777	.502
eb8 ↔ eb7	211.834	.769	eb8 ↔ eb7	237.551	.676
ebl0↔ eb9	630.407	.945	ebl0↔ eb9	607.798	.646
Reg. weights			**Reg. weights**		
b8 ← b7	104.711	.299	b7 ← b8	101.992	.273
b9 ← bl0	356.275	.552	b8 ← b7	102.736	.285
bl0← b9	350.801	.598	b9 ← bl0	322.492	.522
			bl0← b9	322.705	.547

Table 6.8: MI of Importance measurement model.

Pretest			*Posttest*		
Covariances	**M.I.**	**Par chg.**	**Covariances**	**M.I.**	**Par chg.**
ec2 ↔ ecl	185.815	.211	ec2 ↔ ecl	279.158	.206
ec6 ↔ ec5	141.265	.323	ec6 ↔ ec5	257.823	.473
ec8 ↔ ec7	257.733	.548	ec7 ↔ ec6	134.987	.414
ecl0↔ ec9	648.701	.691	ec8 ↔ ec7	326.227	.617
			ecl0 ↔ ec9	623.912	.516
Reg. weights			**Reg. weights**		
c8 ← c7	102.480	.261	c5 ← c6	105.593	.212
c9 ← cl0	306.767	.470	c7 ← c8	138.723	.312
cl0← c9	292.686	.491	c8 ← c7	133.997	.309
			c9 ← cl0	265.029	.421
			cl0 ← c9	254.730	.436

The MI suggested by the software are in Tables 6.6 to 6.8. In the tables, labels for the error terms and the indicators follow the same labels as in the models in Figures 6.5a to 6.5c. The indicators are labeled as **al** to **al0**, **bl** to **bl0**, and **cl** to **cl0** for knowledge, application, and importance respectively. Their relevant error terms are indicated by the letter **e** before the name of the indicator, for example, **eb3** is the error term for indicator **b3**, which in turn is the third indicator for application.

There are two parts of the modification suggested. The first part is concerned with the error terms, where some of them are suggested to covary. For ease of examination, the summary of this part is produced and presented in Table 6.9. The second part is concerned with the direct effects among the indicators, and again this is produced in Table 6.10 on page 105.

The summary of the suggested modifications to the models (Table 6.9 and 6.10) show that the pairs of error terms which are suggested to covary are the same for

Table 6.9: Suggested correlated error terms of the indicators of the LQ measurement models.

	Knowledge	*Application*	*Importance*
Time 1	1 and 2	1 and 2	1 and 2
	5 and 6	5 and 6	5 and 6
	7 and 8	7 and 8	7 and 8
	9 and 10	9 and 10	9 and 10
Time 2	1 and 2	1 and 2	1 and 2
	5 and 6	5 and 6	5 and 6
	7 and 8	7 and 8	6 and 7
	9 and 10	9 and 10	7 and 8
			9 and 10

Table 6.10: Suggested direct effects between indicators of the LQ measurement models.

	Knowledge	*Application*	*Importance*
Time 1	2 to 1	7 to 8	7 to 8
	1 to 2	10 to 9	10 to 9
	9 to 10	9 to 10	9 to 10
	10 to 9		
Time 2	2 to 1	8 to 7	6 to 5
	1 to 2	7 to 8	8 to 7
	8 to 7	10 to 9	7 to 8
	7 to 8	9 to 10	10 to 9
	10 to 9		9 to 10
	9 to 10		

all three subscales across the two time points. The pairs are **1 and 2**, **5 and 6**, **7 and 8**, and **9 and 10**. The only exception is an additional pair between error terms number **6 and 7** in importance time 2.

For the direct effects between the indicators, most of the suggestion are two-ways, meaning that the indicators are suggested to affect one another in both directions. In the models this is represented by a two-way arrow. There are exceptions however, of two instances involving the direct effect of indicators **7 to 8**. All others involve suggestion for both directions between the pairs of indicators. A common pair which is suggested in all cases is between indicators **9 and 10**.

A further examination on the variables shows that the pairs are highly correlated, especially in comparison to the rest of the variables. Values of the correlation are presented in Tables 6.11 to 6.16 on pages 106 to 108.

Inspection of the values in the tables confirms that most of the suggested correlated error terms and suggested direct effects are related to indicators with high correlation coefficients between them. However, the reverse is not necessary true. For example correlations between item 1 and item 3, and item 1 and item 4 of importance pretest are 0.71 and 0.73 respectively, but these pairs are not listed either as suggested error terms or as suggested direct effects.

Table 6.11: Correlation coefficients of the pretest Knowledge indicators.

Indicators	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	**0.70**	1.00								
3	0.25	0.27	1.00							
4	0.42	0.54	0.38	1.00						
5	0.48	0.45	0.31	0.51	1.00					
6	0.44	0.37	0.23	0.32	**0.70**	1.00				
7	0.42	0.30	0.23	0.28	0.54	0.57	1.00			
8	0.40	0.33	0.29	0.32	0.52	0.52	**0.74**	1.00		
9	0.43	0.38	0.29	0.38	0.41	0.35	0.48	0.45	1.00	
10	0.43	0.38	0.27	0.35	0.44	0.39	0.48	0.43	**0.93**	1.00

Table 6.12: Correlation coefficients of the posttest Knowledge indicators.

Indicators	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	**0.78**	1.00								
3	0.33	0.42	1.00							
4	0.50	0.56	0.52	1.00						
5	0.41	0.40	0.27	0.50	1.00					
6	0.41	0.39	0.28	0.34	**0.72**	1.00				
7	0.43	0.38	0.33	0.32	0.48	0.65	1.00			
8	0.39	0.36	0.30	0.32	0.56	0.65	**0.77**	1.00		
9	0.42	0.40	0.36	0.50	0.37	0.29	0.37	0.36	1.00	
10	0.43	0.42	0.37	0.50	0.37	0.29	0.38	0.36	**0.89**	1.00

Specification of correlated error terms gives one way of *multidimensional measurement* (Kline, p. 168); the other is letting indicators load on more than one factor. A measurement error correlation reflects the assumption that the two corre-

Table 6.13: Correlation coefficients of the pretest Application indicators.

Indicators	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	**0.75**	1.00								
3	0.51	0.56	1.00							
4	0.50	0.58	0.60	1.00						
5	0.63	0.53	0.53	0.58	1.00					
6	0.59	0.46	0.41	0.44	**0.77**	1.00				
7	0.52	0.40	0.39	0.36	0.52	0.57	1.00			
8	0.46	0.38	0.41	0.40	0.57	0.60	**0.75**	1.00		
9	0.45	0.44	0.38	0.40	0.43	0.37	0.42	0.41	1.00	
10	0.47	0.44	0.38	0.37	0.42	0.38	0.42	0.41	**0.94**	1.00

Table 6.14: Correlation coefficients of the posttest Application indicators.

Indicators	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	**0.81**	1.00								
3	0.50	0.58	1.00							
4	0.50	0.58	0.60	1.00						
5	0.60	0.55	0.45	0.59	1.00					
6	0.58	0.48	0.37	0.45	**0.80**	1.00				
7	0.56	0.46	0.40	0.35	0.56	0.64	1.00			
8	0.48	0.43	0.43	0.39	0.61	0.64	**0.79**	1.00		
9	0.47	0.46	0.52	0.49	0.46	0.37	0.43	0.44	1.00	
10	0.47	0.48	0.48	0.47	0.46	0.38	0.43	0.45	**0.94**	1.00

Table 6.15: Correlation coefficients of the pretest Importance indicators.

Indicators	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	**0.83**	1.00								
3	0.61	0.67	1.00							
4	0.69	**0.75**	**0.73**	1.00						
5	**0.70**	0.64	0.57	0.65	1.00					
6	0.63	0.54	0.45	0.52	**0.80**	1.00				
7	0.62	0.51	0.48	0.51	0.64	**0.71**	1.00			
8	0.58	0.54	0.52	0.53	0.65	**0.73**	**0.82**	1.00		
9	0.55	0.53	0.53	0.53	0.58	0.48	0.50	0.51	1.00	
10	0.54	0.50	0.52	0.50	0.55	0.47	0.47	0.48	**0.94**	1.00

Table 6.16: Correlation coefficients of the posttest Importance indicators.

Indicators	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	**0.88**	1.00								
3	**0.71**	**0.73**	1.00							
4	**0.73**	**0.76**	**0.73**	1.00						
5	0.67	0.61	0.55	0.65	1.00					
6	0.60	0.55	0.47	0.52	**0.83**	1.00				
7	0.62	0.54	0.51	0.48	0.65	**0.73**	1.00			
8	0.58	0.52	0.49	0.48	0.68	**0.71**	**0.85**	1.00		
9	0.56	0.56	0.60	0.58	0.57	0.48	0.50	0.53	1.00	
10	0.54	0.54	0.59	0.58	0.56	0.47	0.49	0.53	**0.95**	1.00

sponding indicators measure something in common that is not explicitly represented in the model. However, modifying a model by having correlated measurement errors also has implications for the identification of the model. The number of parameters would be increased, reducing parsimony, a sought after characteristic of a structural equation model.

Observations on the values of the MI also reveal a common pattern. In the case of MI for covariances, the largest MI always relate to error terms for items 9 and 10. In the questionnaire, these two items relate specifically to the **knowledge** and **skill**, respectively, of the training programme. In the first model (knowledge), these questions ask the participants to evaluate their level of knowledge and level of skills, with regard to the training programme they are attending. Similarly for the second model (application), the participants are asked to evaluate the level of application of the 'knowledge and skills', of the area focused on by the training. For the third model (importance), the questions ask about the importance of the 'knowledge and skills' in the focus area.

The SEM results that suggest there should be a covariance between these two terms indicate that the two indicators are measuring something in common [54]. This could easily be understood as the two questions are related specifically to the training programme attended, while all the other eight questions in the model are asking about other subject areas in general.

The same error terms are also related to another suggestion by the MI for modification. In the case of regression weights, the highest modification indices values are linked to direct effects between indicators 9 and 10. The MI have suggested that there is a direct link from indicator 9 to indicator 10, and vice versa. This suggestion is not in line with the theory underlying the model. It could be inferred however that these two indicators are strongly related and not co-independent as hypothesized. Removing one of these two items from analysis could be one way of making the models fit the data better.

Only suggestions with modification indices greater than 100 are considered in order to minimise the number of changes to the models. A greater reduction in χ^2 values should be possible if all suggestions are taken into consideration, but that would cause the model to be less parsimonious.

6.3.2 Modified LQ Model

After taking into consideration all the suggested modifications, the new models of knowledge, application, and importance should leave out one item from each of the highly correlated pairs of the indicator variables. As discussed in Section 6.3.1, the highly correlated pairs of indicators are **1 and 2**, **5 and 6**, **7 and 8**, and **9 and 10**. One way of taking out the indicators is by looking at the values of Cronbach's alpha when the item is omitted. The item which reduces alpha more is taken out.

For the knowledge subscale, indicators taken out are 2, 6, 8 and 9. For the application subscale items 2, 6, 7 and 10 are taken out. Items 2, 6, 7 and 10 are taken out from the importance subscale. This leaves only six indicators per factor or subscale, as shown in Figure 6.6, and explained in details in Table 6.17.

In this modified model of Learning (Figure 6.6), the first factor loading from each latent variable is fixed to unity to scale the variables. Similarly, the loading from Learning to knowledge is also fixed to 1. The structural equation of the new LQ model is then:

$$\eta = \beta\eta + \Gamma\xi + \zeta,$$

where

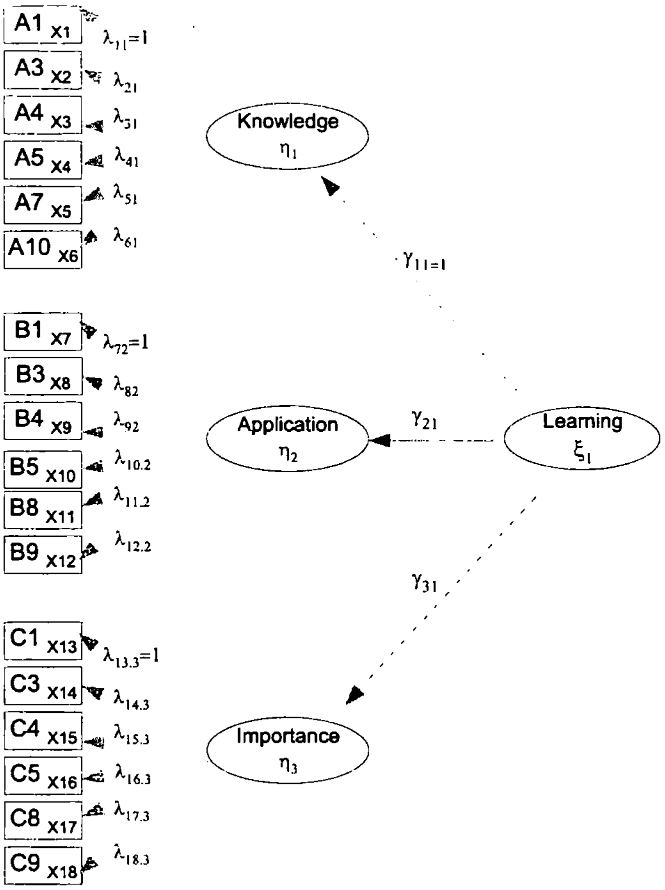


Figure 6.6: The modified Learning model with six indicators per factor.

$$\eta_{(3\times1)} = \begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\end{bmatrix}, \quad \Gamma_{(3\times1)} = \begin{bmatrix}1\\ \gamma_{21}\\ \gamma_{31}\end{bmatrix}, \quad \xi_{(1\times1)} = \begin{bmatrix}\xi_1\end{bmatrix}, \quad \text{and} \quad \zeta_{(3\times1)} = \begin{bmatrix}\zeta_1\\ \zeta_2\\ \zeta_3\end{bmatrix}.$$

In the structural model, the β matrix is zero because the endogenous variables, which are the three subscales (knowledge, application and importance) are independent of each other. They have a common higher factor (Learning), thus their associations are assumed to exist only through the higher factor and not directly among them, and thus not analysed [54]. The equation and the matrices of the equation are as follows:

$$y = \Lambda_y \eta + \varepsilon$$

where

Table 6.17: Variables, indicators, and subject areas of the modified LQ model.

Variable	Indicator	Subject area
Knowledge subscale		
Y1	1	Economic Management
Y2	3	ICT
Y3	4	Human Resource and Organisation
Y4	5	Social and Infrastructure Planning
Y5	7	International Relation
Y6	10	Specific skills (of the programme)
Application subscale		
Y7	1	Economic Management
Y8	3	ICT
Y9	4	Human Resource and Organisation
Y10	5	Social and Infrastructure Planning
Y11	8	Defense and National Security
Y12	9	Specific knowledge (of the programme)
Importance subscale		
Y13	1	Economic Management
Y14	3	ICT
Y15	4	Human Resource and Organisation
Y16	5	Social and Infrastructure Planning
Y17	8	Defense and National Security
Y18	9	Specific knowledge (of the programme)

$$y_{(18\times1)} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{18} \end{bmatrix}, \quad \Lambda_{(18\times3)} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{2.1} & 0 & 0 \\ \lambda_{3.1} & 0 & 0 \\ \lambda_{4.1} & 0 & 0 \\ \lambda_{5.1} & 0 & 0 \\ \lambda_{6.1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{8.2} & 0 \\ 0 & \lambda_{9.2} & 0 \\ 0 & \lambda_{10.2} & 0 \\ 0 & \lambda_{11.2} & 0 \\ 0 & \lambda_{12.2} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{14.3} \\ 0 & 0 & \lambda_{15.3} \\ 0 & 0 & \lambda_{16.3} \\ 0 & 0 & \lambda_{17.3} \\ 0 & 0 & \lambda_{18.3} \end{bmatrix}, \quad \eta_{(3\times1)} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}, \quad \varepsilon_{(18\times1)} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{18} \end{bmatrix}.$$

The y matrix represents the eighteen indicator items. The Λ matrix represents loadings of the three factors (knowledge, application, and importance) on the eighteen items. Each factor is indicated by six items, the first of which is fixed to 1 for identification of the model. The η matrix is the matrix of the three latent factors. The ε matrix represents the errors of the eighteen indicators.

6.4 Simultaneous Factor Analysis for Measurement Models.

In this section, the three models of the Learning subscales are tested whether the same models hold for both time points. In other words, we are testing whether the model of pretest knowledge is the same as that of the posttest knowledge, and similarly for the models of application and importance. What we are not testing in this section is whether the models are correct; we are only testing whether they hold for both populations. The results of the tests are as in Table 6.18.

Table 6.18: Results of simultaneous FA on three Learning subscales.

Scale	**Knowledge**	**Application**	**Importance**
Model χ^2	3724.616	4303.480	4926.449
Prob level	.000	.000	.000

This particular simultaneous factor analysis tests for common models across time points. The null hypotheses for this particular test is that the measurement models of the three LQ subscales are the same for pretest and postest. In these tests the null hypotheses are rejected. Results indicate that all three measurement models do not hold for both time 1 and time 2. All three Learning subscales do not have common models for pretest and posttest. In other words, pretest model is not the same as posttest model, for all three LQ subscales. Like what was mentioned above, this test does not in any way test for model fit.

6.5 Learning Models Based on Exploratory Factor Analysis

The structural model of Learning and the measurement models of its three subscales have been shown not to fit the data. This implies that the suggested models are not correct for the data.

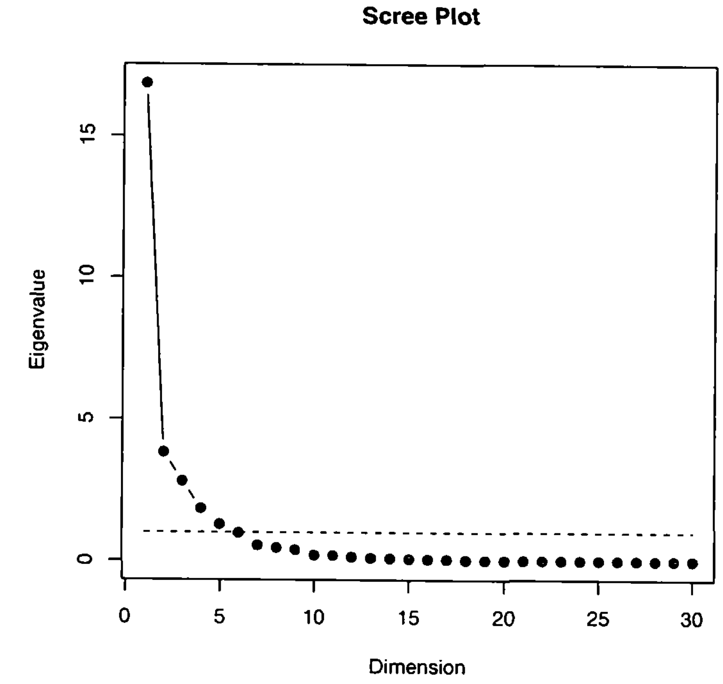


Figure 6.7: Items of pretest Learning.

A scree plot derived from factor analysis on the items of pretest Learning is as in Figure 6.7. There are five factors that have eigenvalues greater than 1, but only the first two seem to reflect underlying factors. Maximum likelihood factor analysis using varimax rotation was carried out on the data, producing loadings as in Table 6.19. Horizontal lines in the table separate the items according to the factors there are supposed to measure in the proposed model.

The first ten indicators have highest loadings on a common factor (Factor 2). This is in agreement with the proposed model of Learning where the first ten items are indicators for a factor, namely knowledge. The following twenty indicator items

Table 6.19: Loadings of pretest Learning by EFA

Number	Items	Factor 1	Factor 2	Factor 3
1	x22		**0.554**	
2	x23	0.137	**0.478**	
3	x24	0.105	**0.345**	
4	x25	0.153	**0.429**	0.106
5	x26	0.109	**0.683**	
6	x27	0.130	**0.671**	
7	x28		**0.742**	
8	x29		**0.714**	
9	x30		**0.719**	0.240
10	x31		**0.729**	0.236
11	x33	**0.676**	0.210	
12	x34	**0.643**		0.141
13	x35	**0.605**		0.150
14	x36	**0.647**		0.130
15	x37	**0.716**	0.270	
16	x38	**0.678**	0.294	
17	x39	**0.539**	0.380	
18	x40	**0.571**	0.375	
19	x41	0.372	0.271	**0.600**
20	x42	0.347	0.298	**0.617**
21	x44	**0.761**		0.275
22	x45	**0.737**		0.274
23	x46	**0.621**		0.330
24	x47	**0.698**		0.290
25	x48	**0.787**	0.117	0.271
26	x49	**0.768**	0.170	0.158
27	x50	**0.686**	0.161	0.226
28	x51	**0.717**	0.164	0.206
29	x52	0.425	0.133	**0.874**
30	x53	0.409	0.135	**0.879**

have highest loadings on another single factor (Factor 1), except for four items, namely items number 19, 20, 29, and 30 (Items x41, x42, x52, and x53 respectively). These four items that load highly on a separate factor (Factor 3) are two pairs of the **focus items**, which are indicator items that represent the *skills* and the *knowledge* specific to the training programme.

The other sixteen items (x33 to x40, and x44 to x51) are found to have their highest loadings on Factor 1 only, even though in the originally proposed model they are supposed to indicate two distinct factors.

While the four items have their highest loadings on Factor 3, they also load on Factor 1 quite highly. This suggests that Factor 3 could be dropped altogether, leaving just Factor 1 and Factor 2. This would agree with the scree plot (Figure 6.7) where we see that there are only two main factors. Furthermore, the two pairs of focus items have been shown to be highly correlated (Please see Section 6.3.1), and for further analyses one item from each pair is suggested to be omitted.

As a summary, indicator items for the application and importance factors are grouped together, but the four focus items for these two factors are clearly loaded to another factor. Generally, this analysis suggests that there are only two dominant factors. The first one is indicated by the ten knowledge items, and the other one is indicated by all application and importance items, except the focus items. The four focus items of application and importance are grouped in a separate factor 3, but because the items also load quite highly on Factor 1, Factor 3 could be dropped altogether.

6.6 Conclusions of Analysis

Confirmatory factor analysis on the measurement models of CEQ showed mixed results, but there are more indices that show the models fit the data than those that show otherwise. Two structural models are subsequently tested. Both models are found to fit the data very loosely. The fits are not very good, but it could also be said that they are not very bad either.

The GHQ is a one-level one-factor model, therefore only the measurement model is tested. The result is not very promising, with none of the fit indices showing good fit. However, the scree plot does show a single dominant factor. The modification indices suggest high correlation between some items.

The measurement models of the LQ are found not to fit the data at all. The modification indices suggest a common pattern of highly correlated items across all three subscales. The simultaneous factor analysis which compares the measurement models of the pretest and posttest data shows that the three Learning subscales do not have the same models at both time points.

The exploratory factor analysis on Learning however clearly indicates two dominant factors. The first factor has the loadings of all application and importance items, except the focus items. The second factor has the loadings of all knowledge items, just as intended in the proposed model.

Chapter 7

Results 3 : Dimension Reduction via Principal Variables Analysis (PVA)

In this chapter we try to reduce the dimension of each of the datasets. This is done at the measurement levels of each scales and subscales. Reducing the dimension of the data using this approach will identify the important variables which are called the *principal variables*, and consequently will identify the redundant and uninformative variables. The method to be employed is *principal variables analysis* (PVA) as proposed by Cumming and Wooff [23], and Cumming [22]. As pointed out by the authors, the particular advantage of this method is that once the principal variables are identified, the remaining variables could be discarded.

The PVA works by calculating a value which is called the h statistics for all variables [22]. The value of h_j is the mean squared covariance between variable j and other variables, and it indicates the amount of contribution of that variable to the overall variability in the dataset. The variable with the largest h value provides the greatest variability of all variables and is taken to be the most desirable to be retained in the scale.

The analysis as performed on these data produces two graphical outputs. First is the scree type plot of the percentage of trace. This plot shows the percentage of variation explained as the variables are selected. The variable which explains the most variation is always extracted first. The second variable extracted by the PVA is the one that explains the most variation after the first one has been selected. Hence, in the scree plot, as more variables are selected through the extraction process, the total percentage of explained variation increases. If each of the variables explains equal amount of variation in the dataset, then the plot shows a straight diagonal line. Otherwise, the plot shows a convex curve that corresponds to the differences in the amount of variation explained by the variables.

7.1 PVA on the LQ

The LQ used in this study is developed specifically for this purpose. The main latent factor, Learning, is measured by three sub-factors, namely knowledge, application, and importance. Each of these sub-factors, which is also latent, is measured by ten items. The ten items are made up of two parts; each of the first eight refers to one subject area, and the remaining two refers to the *skills* and the *knowledge* targetted by the training programme attended by the respondents. Thus the ten items are repeated for all three sub-factors. Other previous analyses have cast a doubt on whether all items are needed for further analyses, hence the justification for dimension reduction.

7.1.1 PVA on the Knowledge Subscale

The following table shows the order of extraction, the names of the variables, the numbers of the variables in the scale and the percentage of trace of each of the variables.

Order :	1	2	3	4	5	6	7	8	9	10
Variables :	v26	v30	v23	v24	v28	v25	v27	v22	v29	v31
Number :	5	9	2	3	7	4	6	1	8	10
Trace % :	32%	19%	11%	9%	9%	6%	5%	4%	4%	1%

The plot of the percentage of trace (Figure 7.1) shows a slight curve, indicating a slow decrease in the amount of variation explained by the items of knowledge. The first variable extracted, number 5 (v26), explains about 32% of total variation in

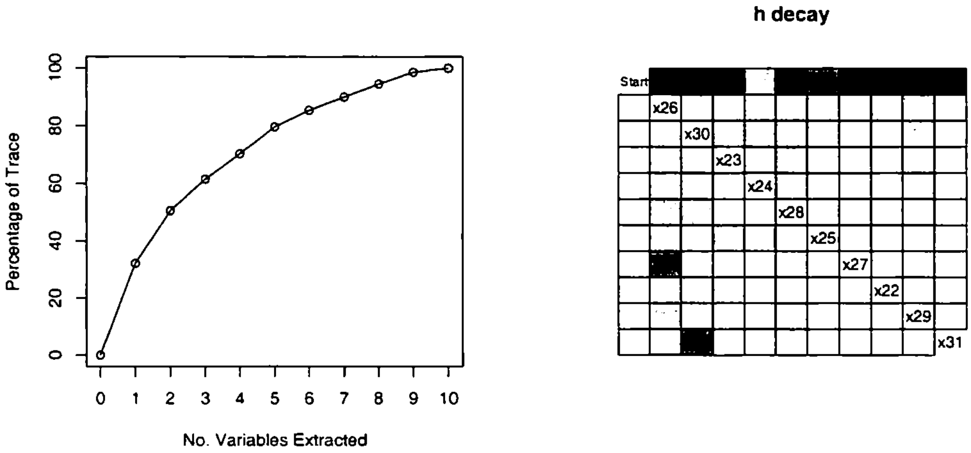


Figure 7.1: Output of Principal Variables on Knowledge.

knowledge. The second variable (v30 - number 9) explains approximately 19% of total variation after variable v26 has been extracted. The third variable (number 2) explains a further 11% of total variation after the first two have been extracted. With the first three variables extracted, over 62% of variation in the data is explained. With four variables the figure is about 71%, and with five variables out of ten it is about 80%. Variables v25,...,v29 contribute almost the same amount of variation explained as the fifth variable (as indicated by the almost straight line in the scree plot), while the contribution of the last one, v31 is negligible. Table 7.1 on page 120 shows the details of PVA for knowledge.

7.1.2 PVA on the Application Subscale.

The order of variables extracted for the application subscale are in the following table.

Order :	1	2	3	4	5	6	7	8	9	10
Variables :	v37	v42	v34	v39	v35	v36	v40	v38	v33	v41
Number :	5	10	2	7	3	4	8	6	1	9
Trace % :	38%	18%	11%	10%	6%	5%	4%	4%	3%	1%

The curve of the percentage of trace plot in Figure 7.2 is steep for the first few variables and starts to decrease after that. The first few items explain much of the

Table 7.1: Results of PVA on Knowledge.

Variables :	v26	v30	v23	v24	v28	v25	v27	v22	v29	v31

Knowledge factor scores.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
v22	2.846	1.021	0.708	0.234	0.234	0.210	0.207	0.202	0.000	0.000
v23	2.687	1.071	0.825	0.000	0.000	0.000	0.000	0.000	0.000	0.000
v24	1.746	0.947	0.792	0.751	0.000	0.000	0.000	0.000	0.000	0.000
v25	2.434	0.771	0.614	0.392	0.343	0.333	0.000	0.000	0.000	0.000
v26	3.214	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
v27	2.826	0.347	0.321	0.314	0.314	0.217	0.210	0.000	0.000	0.000
v28	3.003	0.905	0.588	0.588	0.588	0.000	0.000	0.000	0.000	0.000
v29	2.950	0.902	0.604	0.597	0.585	0.173	0.172	0.170	0.170	0.000
v30	3.166	1.524	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
v31	3.174	1.438	0.021	0.021	0.020	0.020	0.019	0.018	0.018	0.017

Values

	h_k	$h\_k\ w\_k$	$\|\|R\_22.1\|\|^2$	$\|\|R\_11\|\|^2$
Initial	0.000000	0.000000	28.043292	0.000000
x26	3.213712	3.213712	8.926320	1.000000
x30	1.524386	1.524386	4.473788	2.341752
x23	0.825438	0.825438	2.898125	4.041280
x24	0.751289	0.751289	2.084522	5.580852
x28	0.588053	0.588053	0.951990	7.893065
x25	0.333298	0.333298	0.608131	10.749322
x27	0.209861	0.209861	0.390184	14.180819
x22	0.201769	0.201769	0.188377	18.181515
x29	0.170332	0.170332	0.017439	22.695585
x31	0.017439	0.017439	0.000000	28.043292

variation of the application subscale, while the others explain less. Variable number 5 is the first to be extracted and it explains about 38% of the variation. The second variable extracted (number 10) explains a further 14% of total variation, and the two following that (number 2 and 7) contribute to about 11% and 10% respectively. The first four variables explain about 77% of total variation in application. Details of the principal variables extraction are presented in Table 7.2 on page 122.

7.1.3 PVA on the Importance Subscale.

Variables and the order they were extracted are presented in the following table.

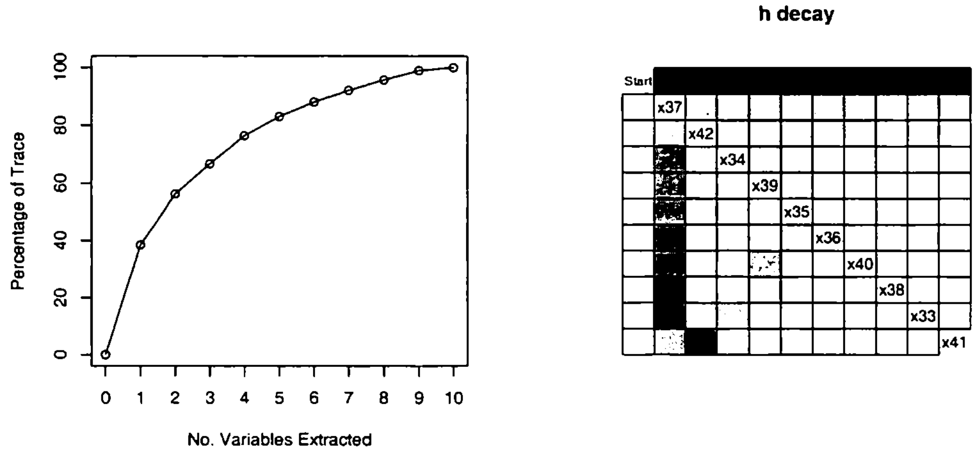


Figure 7.2: Output of Principal Variables on Application

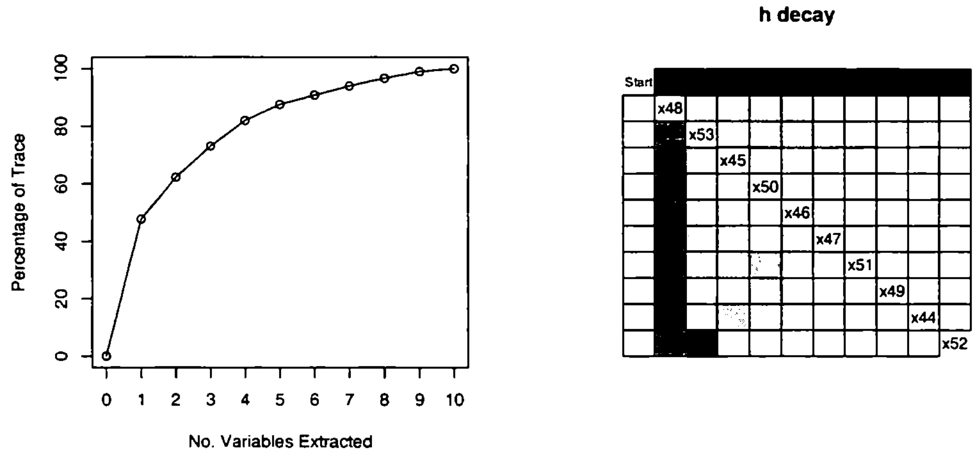


Figure 7.3: Output of Principal Variables on Importance

Order :	1	2	3	4	5	6	7	8	9	10
Variables :	v48	v53	v45	v50	v46	v47	v51	v49	v44	v52
Number :	5	10	2	7	3	4	8	6	1	9
Trace % :	47%	15%	11%	10%	4%	3%	3%	3%	3%	1%

Figure 7.3 shows that the first variable (number 5) extracted for the importance subscale explains approximately 47% of total variation. The second variable extracted (number 10) explains a further 14% of variation, and the third (number 2) explained a further 11% after the first two. The first three variables contribute to

Table 7.2: Results of PVA on Application.

Variables :	v37	v42	v34	v39	v35	v36	v40	v38	v33	v41

Application factor scores

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
v33	3.719	0.713	0.500	0.139	0.115	0.115	0.112	0.111	0.104	0.000
v34	3.401	0.968	0.698	0.000	0.000	0.000	0.000	0.000	0.000	0.000
v35	2.977	0.793	0.641	0.413	0.403	0.000	0.000	0.000	0.000	0.000
v36	3.060	0.658	0.549	0.347	0.347	0.251	0.000	0.000	0.000	0.000
v37	3.837	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
v38	3.476	0.244	0.224	0.215	0.145	0.141	0.140	0.129	0.000	0.000
v39	3.216	0.919	0.686	0.651	0.000	0.000	0.000	0.000	0.000	0.000
v40	3.262	0.784	0.617	0.603	0.162	0.159	0.158	0.000	0.000	0.000
v41	3.262	1.440	0.013	0.013	0.013	0.013	0.012	0.012	0.011	0.011
v42	3.253	1.465	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Values

	h_k	$h\_k\ w\_k$	$\|\|R\_22.1\|\|^2$	$\|\|R\_11\|\|^2$
Initial	0.000000	0.000000	33.463295	0.000000
v37	3.837044	3.837044	7.984835	1.000000
v42	1.465199	1.465199	3.926772	2.353191
v34	0.697743	0.697743	2.380109	4.294617
v39	0.650738	0.650738	1.184309	6.515199
v35	0.402967	0.402967	0.678811	9.276241
v36	0.251314	0.251314	0.421709	12.880067
v40	0.158129	0.158129	0.251902	16.925285
v38	0.129406	0.129406	0.115530	21.900773
v33	0.104324	0.104324	0.010812	27.940223
v41	0.010812	0.010812	0.000000	33.463295

about 74% of the variation in the importance subscale. Further details are presented in Table 7.3 on page 123.

7.2 Summary of PVA Analyses on the LQ subscales

Table 7.4 shows the items of each factor according to the order they were selected for all three Learning subscales. The cumulative percentage of total variation explained as the items are selected is also presented. In the second part of the table is a list of all the items and the subject areas they refer to.

Table 7.3: Results of PVA on Importance.

Variables :	v48	v53	v45	v50	v46	v47	v51	v49	v44	v52

Importance factor scores

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
v44	4.754	0.589	0.446	0.077	0.056	0.056	0.056	0.054	0.053	0.000
v45	4.503	0.766	0.602	0.000	0.000	0.000	0.000	0.000	0.000	0.000
v46	3.920	0.831	0.592	0.282	0.273	0.000	0.000	0.000	0.000	0.000
v47	4.358	0.688	0.540	0.180	0.177	0.104	0.000	0.000	0.000	0.000
v48	4.770	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
v49	4.269	0.213	0.207	0.204	0.094	0.090	0.089	0.073	0.000	0.000
v50	4.190	0.651	0.553	0.500	0.000	0.000	0.000	0.000	0.000	0.000
v51	4.283	0.623	0.530	0.469	0.097	0.095	0.095	0.000	0.000	0.000
v52	4.105	0.983	0.014	0.012	0.011	0.011	0.011	0.011	0.010	0.010
v53	3.926	1.015	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Values

	h_k	$h\_k\ w\_k$	$\|\|R\_22.1\|\|^2$	$\|\|R\_11\|\|^2$
Initial	0.000000	0.000000	43.076278	0.000000
v48	4.769622	4.769622	6.359795	1.000000
v53	1.014659	1.014659	3.482206	2.606192
v45	0.601914	0.601914	1.724414	4.941396
v50	0.500275	0.500275	0.707490	7.724754
v46	0.273156	0.273156	0.354983	11.274404
v47	0.103528	0.103528	0.250482	16.345407
v51	0.094567	0.094567	0.137547	21.661067
v49	0.072781	0.072781	0.063508	27.956560
v44	0.053151	0.053151	0.009912	35.866546
v52	0.009912	0.009912	0.000000	43.076278

Table 7.4: Selection order of the items of LQ subscales.

Order	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Knowledge	5	9	2	3	7	4	6	1	8	10
Cum. trace %	32%	51%	62%	71%	80%	86%	91%	95%	99%	100%
Application	5	10	2	7	3	4	8	6	1	9
Cum. trace %	38%	56%	67%	77%	83%	88%	92%	96%	99%	100%
Importance	5	10	2	7	3	4	8	6	1	9
Cum. trace %	47%	62%	73%	83%	87%	90%	93%	96%	99%	100%

Items of the LQ subscales and the subjects they refer to.

Item	*Subject*
1	Economic Management.
2	Financial Management.
3	Information Technology & Communication
4	Human Resource Management.
5	Social & Infrastructure Planning and Administration.
6	Local Government Administration.
7	International Relations.
8	Defense/Security
9	*Knowledge* specific to the training programme.
10	*Skills* specific to the training programme.

Items 9 and 10 in the list of subjects actually refer to the *knowledge* and *skills* specifically targetted by the training programme attended by the study participants. If, for example, the participants attended a course on information technology, then item 9 refers to, among others, their knowledge on how information technology can be used effectively in management, and item 10 refers to, among others, their skills in using the computers and networking. In that sense, these two items are the **focus items**, compared to the first eight items which are more general.

In all three cases, the first item selected is item number 5, which refers to the subject of **Social and Infrastructure Planning and Administration** (SIPA). The second item selected is always a focus item. The focus item of *knowledge* is selected from the knowledge subscale, and the focus item of *skills* is selected from the application and importance subscales. The third variable selected is always number 2, which refers to the subject of **Financial Management**. The fourth and fifth variables selected are either item 3 (**Information Technology and Communication**) or item 7 (**International Relations**). Item 4 (**Human Resource Management**) is selected as the sixth principal variable in all three instances.

This shows that item 5 is always the item with the largest h statistics, ie. mean squared correlation with the other variables [23]. The item with the second largest h statistics is always one of the two focus items, with the other one having the least. Based on the findings of this analysis, as well as previous analyses, items 9 and 10 could be surrogating each other. They have correlation coefficients of 0.928 for knowledge, 0.945 for application and 0.944 for importance.

Other less obvious patterns also exist. At positions 7, 8 and 9 there are items 1 (**Economic Management**), 6 (**Land, Territorial, Regional and Local Government Administration**), and 8 (**Defense and National Security**). The other focus item, the one not selected at the second iteration, is at the last position.

Table 7.4 shows that almost 90% of total variation can be explained by just the first six items. This analysis thus recommends that only six items are needed to capture a large part of the total variation in the dataset. The items suggested to be retained in the subscales are 2, 3, 4, 5, 7, and 9 for knowledge, and items 2, 3, 4, 5, 7, and 10 for application and importance. The other four items in the subscales can be discarded. As indicated in Table 7.4, the percentages of variation explained by the six variables are 86%, 88%, and 90% for knowledge, application, and importance respectively.

7.3 Principal Variables on the GHQ Scale

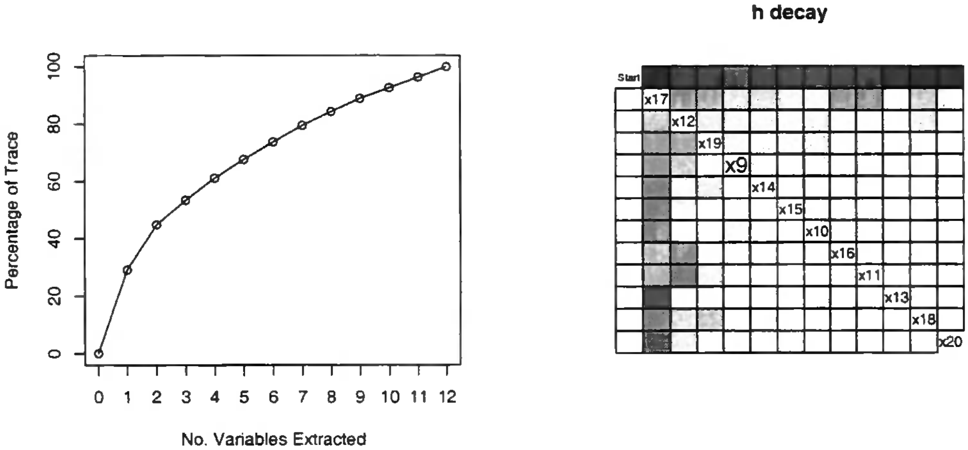


Figure 7.4: Output of Principal Variables on General Health

The following table shows the GHQ variables in the order they are selected by the PVA, with their names and numbers, and the cumulative percentage of trace as the variables are selected.

Order :	1	2	3	4	5	6	7	8	9	10	11	12
Names :	x17	x12	x19	x9	x14	x15	x10	x16	x11	x13	x18	x20
Number :	9	4	11	1	6	7	2	8	3	5	10	12
Cum. trace % :	29%	44%	53%	60%	68%	74%	80%	84%	89%	93%	96%	100%

The curve in the percentage of trace plot (Figure 7.4) is not very steep. This implies that the amount of information explained by each of the variables does not differ very much. The only notable observation is that after the first two variables are selected, the slope seems to be on a straight line. Nevertheless, PVA shows that there is no strong evidence to reduce the dimensionality of the GHQ data.

7.4 Principal Variables on the CEQ Scale

PVA on the CEQ data provides graphical outputs as in Figure 7.5. Variables extracted by the PVA method are presented below, along with the subscales they come from.

Order :	1	2	3	4	5	6	7	8	9	10	11	12
Variables :	y10	y20	y15	y12	y1	y8	y19	y13	y22	y7	y14	y11
Scale :	gs	aw	gt	aa	cg	aa	aa	cg	cg	gt	aw	gs
Order :	13	14	15	16	17	18	19	20	21	22	23	
Variables :	y17	y4	y5	y9	y23	y16	y6	y3	y21	y18	y2	
Scale :	gt	aw	gs	gs	aw	gt	cg	gt	gs	gt	gs	

The curve of percentage trace plot (Figure 7.5) shows a sharper initial increase until the third variable extracted. The line seems to straighten out after that until the last of the 23 variables extracted. The first variable extracted explained about 21% of variation. The first three variables extracted explain about 40% of total variation, and the first eight explain more than 60% of variation. About 80% of total variation is explained by the first fourteen variables.

It can also be observed that the first five variables selected come from five different subscales (GS, AW, GT, AA, and CG). These five variables contain about 50% of the information in the dataset. The increase in the amount of variation explained by the variables is not sharp, thus there is no strong evidence for dimension reduction.

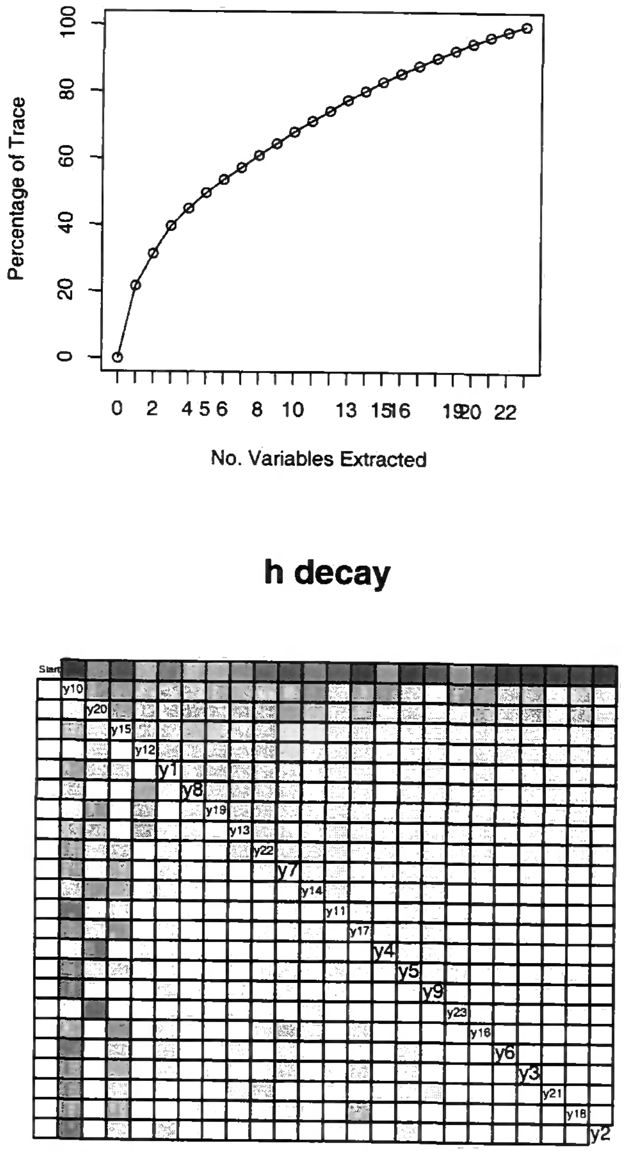


Figure 7.5: Output of Principal Variables on Course Experience

Chapter 8

Results 4 : Survey Respondents

In this chapter we aim to investigate the differences in the scores of the scales and the subscales, between or among the levels of several demographic factors. The factors are (i) *sex*, (ii) *ethnic group*, (iii) *age group*, (iv) *centre*, (v) *service sector*, (vi) *service group*, and (vii) *experience*. Statistical tests of differences are used to examine whether there are statistically significant differences, or whether the differences are just coincidence and can be attributed to chance. Here, two situations are of interest:

First: The scales of CEQ (Section 3.4.1) and GHQ (Section 3.4.2) measure *course experience* factor and *general health* factor respectively. These two scales ideally should be free from the effect of the demographic factors. If this is so, it will be indicated by insignificant differences between or among the levels of the factors. In this situation, insignificant differences will support the idea that the scales are independent of the demographic factors.

Secondly: Based on the literature reviewed, both the CEQ and the GHQ are well established and widely used measurement tools. Assuming that the scales are valid and reliable, it is interesting to examine if the scores differ in any of the demographic factors. For example, it is interesting to find out whether the CEQ differs between the sexes, or among the participants from the different centres.

In this chapter, boxplots are used to present the distributions of the scores. Some of these plots will show some granularity, as a result of the addition of discrete integer items.

8.1 Demographics

A total of seven demographic factors are considered in this section. The factors are introduced below with a brief explanation of each. For three of the factors, namely *sex*, *age*, and *service sector*, latest statistics of 2006 are also provided as comparison. The latest statistics are based on the annual report for 2006 produced by INTAN[1]. Latest statistics for the other factors are not available.

8.1.1 Sex

The number of respondents by *sex* are 424 (56%) males and 333 (44%) females. There are 760 participants altogether, but the sum of these two figures falls short because of non-responses. In other word, there are three respondents who did not indicate their sex. The actual figures of training participants in INTAN for the year 2006 are 49% for male and 51% for female.

8.1.2 Ethnic Group

Ethnicity as a factor has four levels, three of which refer to the main ethnic groups of Malaysia, namely the Malays, the Chinese and the Indians. Respondents from any other smaller ethnic groups are combined into a level called *Others*. The numbers of respondents from each ethnicity level who participated in the study are as follows:

Number of respondents by *ethnic group*.

Ethnicity	*Malay*	*Chinese*	*Indian*	*Others*
Numbers	647	35	49	27
Percentage	85.4%	4.6%	6.5%	3.5%

8.1.3 Age

Age of respondents is grouped into six levels: (1) below 26 years, (2) 26 to 30 years, (3) 31 to 35 years, (4) 36 to 40 years, (5) 41 to 45 years, and (6) above 45 years. These ranges of age are similar to those used by the current programme evaluation of INTAN. The numbers of study participants for each level are in the following

[1] Annual Report 2006; National Institute of Public Administration

table. The differences between the percentages of the study participants and the actual percentages of participants who attended training in 2006 suggest that the distributions of study sample might not be representative of the actual distributions, in terms of age.

Number of respondents by *age group*.

Age	*<26*	*26-30*	*31-35*	*36-40*	*41-45*	*>45*
Numbers	185	357	128	30	31	26
Percentage	24.4%	47.2%	16.9%	4%	4.1%	3.4%
2006 actual	10.6%	21.0%	19.2%	12.0%	24.0%	13.2%

8.1.4 Centre

In INTAN, training programmes are organised and managed by the centres. These centres are known and differentiated by the scope of training they organised. For example, the centre of Financial Management organises only courses related to financial management. Participants who took part in this survey were attending courses organised by six of the nine centres, plus the Institute of Diplomatic and Foreign Relations (IDFR). The IDFR is a training institute under the Ministry of Foreign Affairs and is not part of INTAN, but it cooperates with INTAN in the running of courses related to the diplomatic and foreign policy. The numbers of participants from the six INTAN Centres and IDFR, and the percentages are as follows:

Number of respondents by INTAN Centre and IDFR.

Centres	Numbers	Percentage
1. Management Development (**Mgt**)	26	3.4%
2. Economy Development (**Econ**)	155	20.4%
3. Local Government and District Management (**KTD**)	8	1%
4. Quantitative Centre (**Quant**)	172	22.6%
5. Management Technology (**Imatec**)	155	20.4%
6. Financial Management (**Finance**)	28	3.7%
7. Institute of Diplomatic and Foreign Relation (**IDFR**)	216	28.4%

8.1.5 Service Sector

A respondent's *service sector* refers to the organisational sector where the participants are currently working. It has three levels: Federal, State, and Local. Federal refers to the federal ministries and government departments, while State refers to any of the state governments. Those who work at the local governments and authorities are grouped in Local. The followings are the numbers of participants by service sector. The distributions of the study sample appear to be not far off from the distributions of the actual training participants of 2006, in terms of service sector.

Number of respondent by *service sector.*

Service Sector	*Federal*	*State*	*Local*
Numbers	664	52	41
Percentage	87.7%	6.9%	5.4%
2006 actual	80.0%	8.8%	3.5%

8.1.6 Service Group

A respondent's *service group* refers to the managerial group of the respondents. There are two broad categories, the Professional and Management Group (Prof), and the Supporting Staff (Supp). The Professional and Management groups consists mostly of managers, and most of the Supporting Staff are executives. For this research, the number of Supporting Staff who took part is just 47 (6.2%), compared to 708 (93.8%) from the Professional and Management group.

8.1.7 Experience

In this study, *experience* refers to the number of years the respondents have been working in the Malaysian public sector. It is another factor, other than *age*, whose ranges are based on the current evaluation form. The levels are (1) 5 years or less, (2) 6 to 10 years, (3) 11 to 15 years, and (4) more than 15 years. The distribution of participants according to their experience is as follows:

Number of respondents by range of *experience* (years).

Experience	*5 or less*	*6 to 10*	*11 to 15*	*> 15*
Numbers	638	39	28	55
Percentage	83.9%	5.1%	3.7%	7.3%

8.2 Methodology

Statistical tests of differences are performed to compare the scores of the scales between or among the different levels of the demographic factors. This will indicate whether the differences observed are large enough to suggest actual differences in the population, or are just due to chance. If a difference is found to be statistically significant, then we can say that the scale is associated with the factor.

For factors with two levels (namely *sex* and *service group*), Welch 2-sample t-test is used. The Mann-Whitney or Wilcoxon test is also utilised as the non-parametric alternative for hypothesis testing. For all other factors, oneway analysis of variance (ANOVA) is used, with the Kruskal-Wallis test as the non-parametric alternative. While the standard ANOVA assumes equal variances in the distribution of the scores of the groups, both the Welch 2-sample t-test and the ANOVA used in this study do not assume equal variances. This is because the distributions of all the variables in this study do not usually meet the assumption of equal variances. However, in cases of significant ANOVA tests, the follow up post hoc tests in the form of the Tukey Honestly Significant Difference (Tukey HSD) assumes equal variances. All tests are carried out using the R statistical software (http://www.r-project.org).

When equal variance is not chosen in the Welch 2-sample t-test, the R software estimates the variance separately, and the Welch modification to the degrees of freedom is used [88]. For the oneway ANOVA with this option, R uses an approximate method of Welch (1951) [92]. This method generalises the 2-sample Welch test to the case of many samples [88].

The scores of the variables are computed based on the originally proposed models. Each of the LQ subscales, namely the knowledge, application, and importance, (Chapter 4) is represented by the sum of its 10 item indicators. The score of the GHQ (Section 3.4.2) is the total score of its twelve items. Each of the subscales of the CEQ (Section 3.4.1) is represented by the mean score of its items. The LQ has positive subscales, which means a higher score indicates either a higher level of knowledge, a higher level of application, or a higher level of importance. Similarly the CEQ also has a positive scale, where a higher score indicates a more positive reaction in terms of the factor measured. The GHQ measures general health in a negative scale. A better level of health is indicated by a lower score.

In some cases where individual items are not scored, the total score is zero. To reduce biasness in the score of the scale, cases with zero are considered as missing and omitted from analysis. Consequently, the number of the samples for each test varies depending on the number of cases omitted. This approach of dealing with missing data is known as 'pairwise deletion', or 'available-case analysis' [27]. This is the approach used in this chapter.

Throughout this chapter there are many tests done on the same dataset. In this situation, a Bonferroni correction is usually suggested, where the statistical significance level is adjusted by multiplying it with 1/n, where n is the number of tests ([10], [97]). We will not be using this approach however, instead, we will be quite stringent in the level of significance. This approach produces the same effect as adjusting the significance level. Wordings used to indicate the amount of evidence to reject the null hypotheses are as follows:

P-values	*Conclusion*
≥ 0.01	- **no evidence**
≥ 0.001	- **weak evidence**
≥ 0.0001	- **some evidence**

Some of the variables display evidence of non-normality. However, the sample size is large enough that the t-test should be robust to this deficiency. This is the case for many of the tests in this chapter.

The tests are applied to the factors whose levels are in most cases not equal in size. A statistical test on a large sample will more likely produce a strong evidence against null hypothesis than the same test on a smaller sample. Taking that into consideration, we also need to look at the relative size of the mean differences in making conclusions about the existence of evidence to support significant differences.

8.3 Sex

In this section we explore whether the scales used differ for men and women. There is no priori expectation that any of the scales, the GHQ, CEQ, or LQ, differ between the sexes.

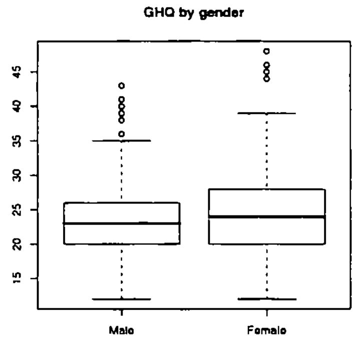


Figure 8.1: GHQ by *sex.*

8.3.1 General Health Questionnaire

The scale of GHQ has twelve items, all of which are measured by a common 4 point scale. Thus minimum score is 12 and the maximum is 48. Overall, the GHQ has a mean of 23.899 and a standard deviation of 5.466. This fits in nicely into the range of a typical score, which is between 23 and 24 ([31]). The distribution of the score is slightly skewed, with an obvious tail to the right of the distributions of both the male and female respondents (Figure 8.1).

In the comparison of the means between the sexes, the hypothesis to be tested is that the means of the general health are the same for male and female respondents. Figure 8.1 shows the dispersion of the general health of female respondents is slightly wider than that of male respondents. The tables below show the summary statistics, namely the number of respondents (n), the mean score ($\overline{x}$), and the standard deviation (s), of the GHQ by sex, followed by the results of the tests of differences.

Summary statistics of the GHQ by *sex.*

	male	*female*
n	424	334
$\overline{x}$	23.55	24.33
s	5.13	5.82

Tests of differences of GHQ by *sex*.

Welch t-test				Mann-Whitney	
df	*t*	*p-value*	95% *CI*	*W*	*p-value*
666.106	-1.92	0.056	(-1.57, 0.02)	65876	0.11

The results of the tests of differences in the table above show no evidence of significantly different means. There seems to be no difference in the general psychological health between male and female participants.

Goldberg had reported in the manual of the General Health Questionnaire that female samples scored higher than male for patient samples [35]. This finding by Goldberg was similar for both types of patients, ie. those in a consulting setting and those at home and not consulting their doctors. The scores were higher among patients in consulting setting, which were more symptomatic than those who were not. However, respondents of the Golberg's study were generally 'ill patients', and should not be compared to the training participants in this study.

8.3.2 Pretest Learning

The LQ has three subscales: knowledge, application and importance. The summary statistics of the pretest scores by *sex* are in the following table.

Summary statistics of pretest LQ subscales by *sex*.

	Sex	*Male*	*Female*
	n	424	334
Knowledge	$\overline{x}$	42.44	41.33
	s	8.58	8.88
Application	$\overline{x}$	50.92	51.67
	s	9.86	10.90
Importance	$\overline{x}$	56.10	56.38
	s	9.58	10.71

Figure 8.2 indicates that the medians and the interquartile ranges of all three LQ subscales are about equal between the sexes. Test statistics for equality of means from both the Welch 2-sample t-test and the Mann-Whitney test are presented below.

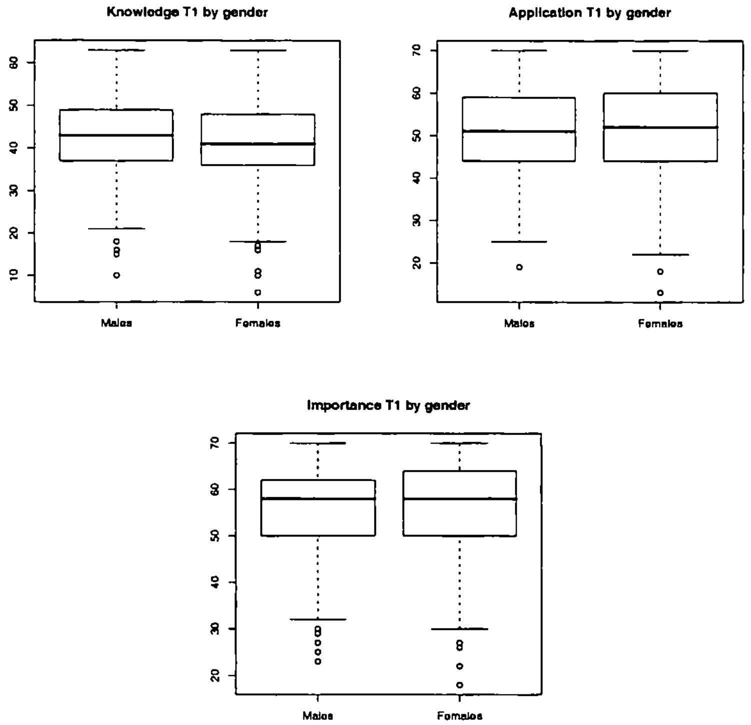


Figure 8.2: Pretest LQ subscales by *sex*.

Test of differences of pretest Learning subscales by *sex*.

Statistics	*Knowledge*	*Application*	*Importance*
df	701.44	678.50	672.23
t	1.74	-0.98	-0.36
p-value	0.08	0.33	0.72
95%CI	(-0.14, 2.37)	(-2.25, 0.75)	(-1.74, 1.20)
W	75946.5	67260	68260
p-value	0.073	0.258	0.466

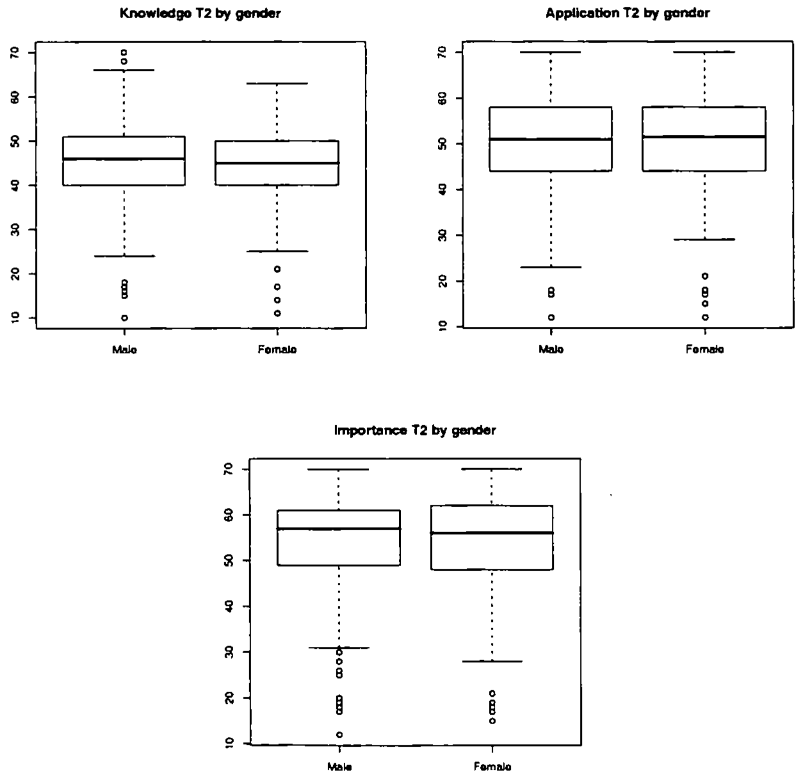
Among the three null hypotheses tested, the one for knowledge has the smallest p-value, but it is still too large to suggest significant difference. Hypothesis of no difference for the other two factors are not rejected at all. Therefore it is concluded that for all pretest LQ subscales, there are no differences between the means of male and female participants. Results from the Mann-Whitney tests re-emphasise the conclusions of the t-tests.

8.3.3 Posttest Learning

The following table shows the summary statistics of the posttest LQ subscales by *sex*. There seem to be very little differences between the means.

Summary statistics of posttest LQ subscales by *sex*.

	Sex	*Male*	*Female*
	n	424	334
Knowledge	$\overline{x}$	45.12	44.72
	s	9.24	8.38
Application	$\overline{x}$	50.32	51.03
	s	10.89	10.30
Importance	$\overline{x}$	54.56	54.70
	s	10.80	10.90


Knowledge T2 by gender

Application T2 by gender

Importance T2 by gender

Figure 8.3: Posttest LQ subscales by *sex*.

Figure 8.3 does not indicate significant differences in the medians or the interquartile ranges. Statistical test results are presented in the table below.

May 31, 2008

Test of differences of posttest LQ subscales by *sex*.

Statistics	*Knowledge*	*Application*	*Importance*
df	738.11	727.04	709.94
t	0.62	-0.92	-0.18
p-value	0.54	0.36	0.86
95%CI	(-0.87, 1.66)	(-2.23, 0.81)	(-1.70, 1.42)
W	72506.5	68094.5	70555.5
p-value	0.417	0.509	0.966

The results of the tests confirm that *sex* is not a factor associated with the posttest Learning. This is true for all three subscales. The average level of Learning of the male respondents after training ended is not different from that of female respondents.

8.3.4 Course Experience Questionnaire (CEQ)

The CEQ consists of five subscales, namely good teaching (GT), clear goals (CG), generic skills (GS), appropriate assessment (AA), and appropriate workload (AW). They are measured by different numbers of indicator items: the GT by 6 items, the CG by 4, the GS by 6, the AA by 3 and the AW by 4 items. All items are measured on five point Likert scales. The table below shows the number of respondents, the means, and the standard deviations of the CEQ subscales by *sex*.

Summary statistics of the CEQ subscales by *sex*.

	Sex	*Male*	*Female*
	n	424	334
GT	$\overline{x}$	3.45	3.44
	s	0.653	0.637
CG	$\overline{x}$	3.45	3.42
	s	0.561	0.547
GS	$\overline{x}$	3.54	3.57
	s	0.634	0.640
AA	$\overline{x}$	3.22	3.21
	s	0.614	0.613
AW	$\overline{x}$	3.10	3.16
	s	0.781	0.766

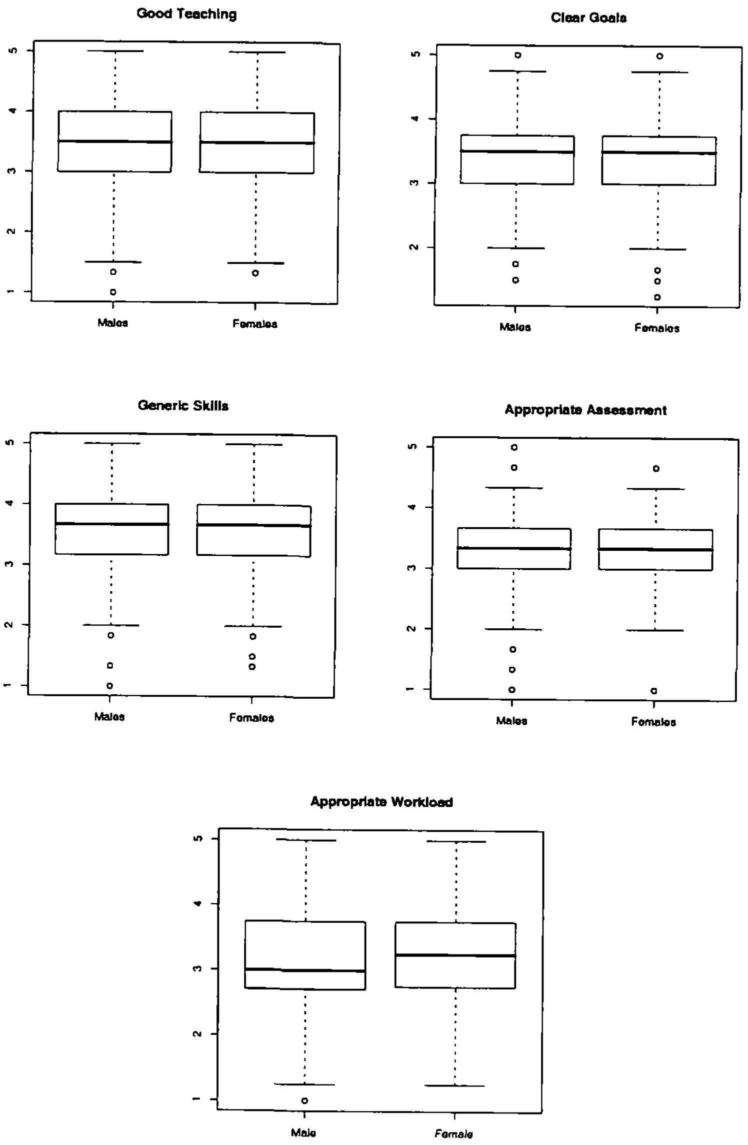


Figure 8.4: CEQ subscales by *sex*.

The boxplots of the distribution are presented in Figure 8.4. The plots suggest that in all cases the variances between the sexes do not differ very much, except for the AW. The medians of the AW also seem to differ between male and female respondents. The results of the tests of differences on the scores of the CEQ subscales by *sex* are presented in the table below.

Tests of differences of the CEQ subscales by *sex*.

	df	*t*	*p-value*	95% CI	*W*	*p-value*
GT	722.70	0.32	0.75	(-0.08, 0.11)	71685	0.769
CG	723.07	0.90	0.37	(-0.04, 0.16)	73109.5	0.437
GS	712.21	-0.59	0.55	(-0.12, 0.06)	69022.5	0.549
AA	715.86	0.34	0.74	(-0.07, 0.10)	71058.5	0.932
AW	721.24	-1.12	0.26	(-0.17, 0.05)	68068	0.358

The results of both the Welch two sample t-test and the Mann-Whitney test (W) indicate that of the five subscales of the CEQ, none rejects the hypothesis that the true means are equal for male and female populations. It can be concluded that male and female respondents have similar training experiences in terms of the five CEQ factors measured by the scales.

8.4 Ethnic Group

In this section, the scores of the scales are compared among the four ethnic groups. One important finding would be nonsignificant differences in the CEQ subscales, as that would suggest participants from any ethnic background have similar training experience.

8.4.1 General Health Questionnaire

Figure 8.5 suggests that participants of Indian ethnicity have a slightly higher general health score than the other three groups. The distributions do not seem to be Normal, but we will assume they are approximately Normal. The means and the standard deviations by ethnicity are presented in the follwing table.

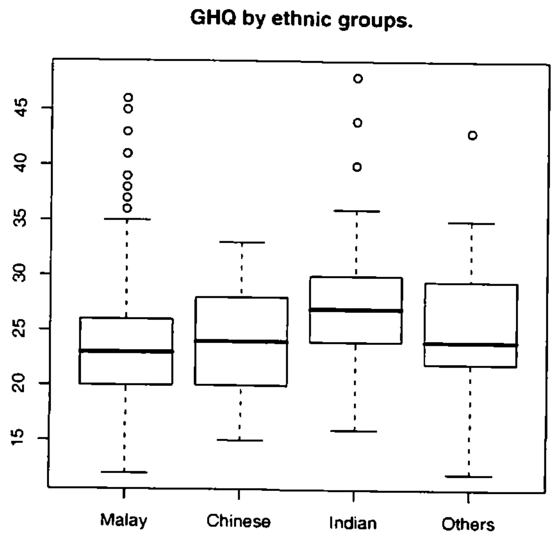


Figure 8.5: GHQ by *ethnic group.*

Summary statistics of the GHQ by *ethnic group.*

	Malay	Chinese	Indian	Others
n	647	35	49	27
$\overline{x}$	23.53	24.51	27.39	25.63
s	5.21	4.76	7.17	6.45

The oneway ANOVA test and the Kruskal-Wallis tests on the data gave these results:

Test of differences of the GHQ by *ethnic group.*

	Statistics	*p-value*
ANOVA	$F_{3,60.85} = 5.53$	0.002
Kruskal Wallis	$W_3 = 20.80$	0.0001

The result shows a weak evidence of significant differences between the means of the GHQ of the ethnic groups. A post-hoc test indicates that the significant difference is between the means of the Indians and the Malays. Participants of Indian ethnicity seem to have the highest score of general health, while those of Malay ethnicity seem to have the lowest. The scores imply that the respondents of Malay ethnicity perceive their general mental health at a better level than the level perceived by their Indian counterparts.

8.4.2 Pretest Learning

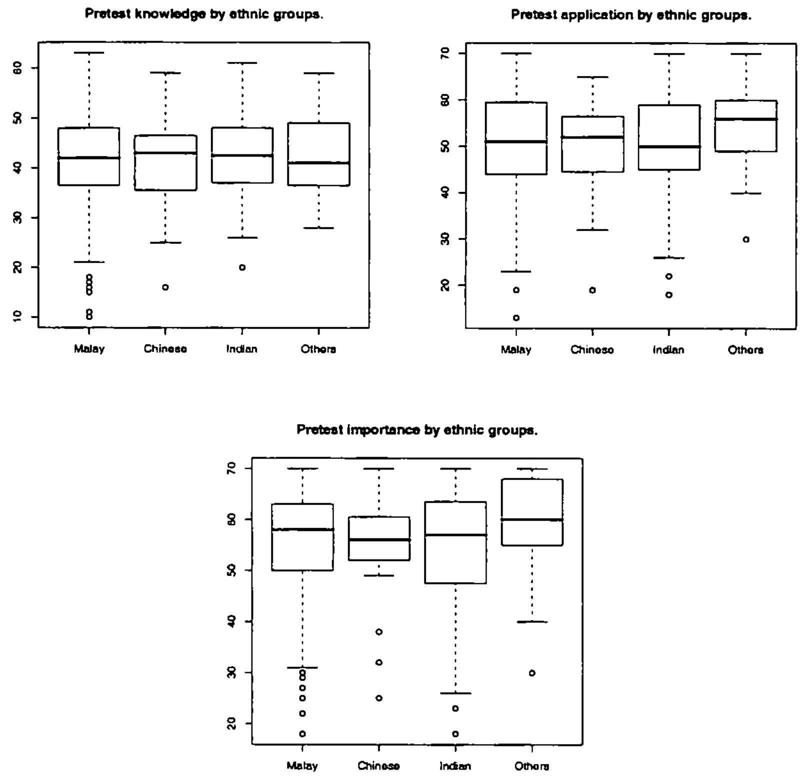


Figure 8.6: Pretest LQ subscales by *ethnic group.*

Figure 8.6 suggests there are differences in the medians, as well as in the interquartile ranges of the ethnic groups. The number of samples, the means and the standard deviations of the three pretest LQ subscales by ethnic groups are presented in the following summary statistics table:

Summary statistics of pretest LQ subscales by ethnic groups.

		Malay	Chinese	Indian	Others
	n	647	35	49	27
Knowledge	$\overline{x}$	41.95	41.17	42.00	42.22
	s	8.81	8.18	8.65	8.46
Application	$\overline{x}$	51.25	49.77	50.73	53.73
	s	10.30	9.99	11.67	9.67
Importance	$\overline{x}$	56.34	55.37	53.97	58.92
	s	9.85	9.68	12.75	10.42

The results of the tests of differences of the LQ subscales among the ethnic groups are presented in the following table:

Test of differences of pretest LQ subscales by *ethnic group*.

	ANOVA			Kruskal		
	df	*F*	*p*	*df*	χ^2	*p*
Knowledge	3, 62.04	0.11	0.95	3	0.27	0.97
Application	3, 60.75	0.84	0.48	3	1.88	0.60
Importance	3, 59.69	1.19	0.32	3	3.34	0.34

The tests suggest no evidence that the participants of different ethnic groups gave different scores to either knowledge, application, or importance. They indicated the same level of perceived knowledge, and they also had the same view as to which subjects are highly used in everyday work and which are not. How they feel about the importance of learning the subjects also seems to be similar.

8.4.3 Posttest Learning

Just like for the pretest data, Figure 8.7 suggests some variation, in both the medians and the interquartile ranges, among the ethnic groups. The number of samples, the means and the standard deviations of the posttest LQ subscales by ethnic groups are presented below. The results of the oneway ANOVA are presented in the table following that.

Summary statistics of posttest LQ subscales by ethnic groups.

		Malay	Chinese	Indian	Others
	n	647	35	49	27
Knowledge	$\overline{x}$	45.07	44.69	42.57	46.96
	s	8.84	7.83	9.87	8.38
Application	$\overline{x}$	50.91	48.66	47.65	52.33
	s	10.44	10.15	12.09	11.91
Importance	$\overline{x}$	54.69	55.26	51.04	58.74
	s	10.59	11.49	13.05	10.03

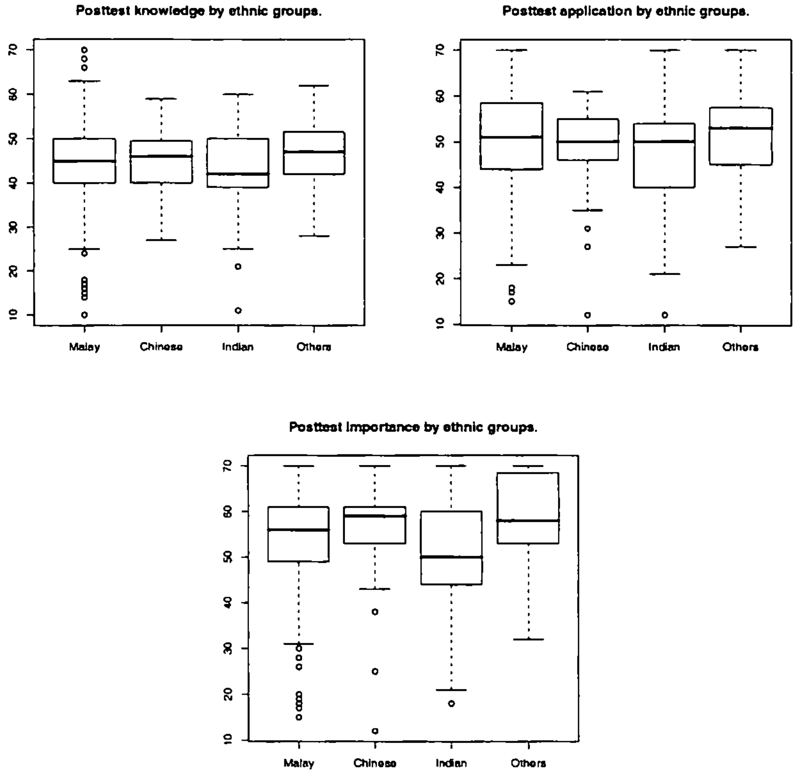


Figure 8.7: Posttest LQ subscales by *ethnic group*.

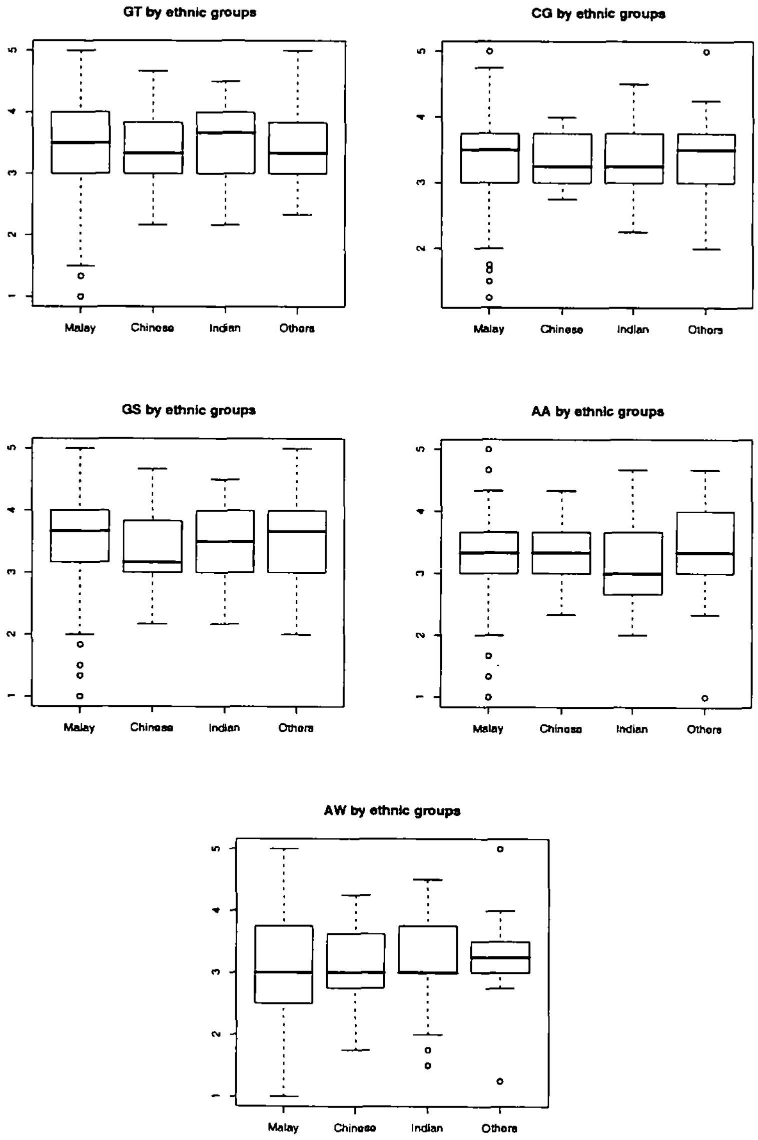
Test of differences of posttest LQ subscales by *ethnic group*.

	ANOVA			Kruskal		
	$df_{num,denom}$	F	p	df	χ^2	p
Knowledge	3, 62.28	1.48	0.23	3	3.96	0.27
Application	3, 61.12	1.73	0.17	3	4.04	0.26
Importance	3, 61.32	2.71	0.05	3	7.81	0.05

The ANOVA test however indicates no evidence of significant differences in the posttest Learning among ethnic groups. The only notable result is a slight evidence of difference in the scores of importance, but it is not strong enough to warrant further examination.

8.4.4 Course Experience Questionnaire

Figure 8.8 indicates some variation in the interquartile ranges of the CEQ subscales among the ethnic groups. The means and the standard deviations of the subscales



GT by ethnic groups

CG by ethnic groups

GS by ethnic groups

AA by ethnic groups

AW by ethnic groups

Figure 8.8: CEQ subscales by *ethnic group*.

by ethnic groups are as follows:

Summary statistics of the CEQ subscales by *ethnic group.*

		Malay	*Chinese*	*Indian*	*Others*
	n	647	35	49	27
GT	$\overline{x}$	3.45	3.40	3.47	3.41
	s	0.66	0.54	0.61	0.60
CG	$\overline{x}$	3.45	3.35	3.37	3.40
	s	0.56	0.38	0.58	0.66
GS	$\overline{x}$	3.57	3.35	3.47	3.54
	s	0.64	0.53	0.58	0.72
AA	$\overline{x}$	3.21	3.28	3.22	3.31
	s	0.60	0.56	0.67	0.78
AW	$\overline{x}$	3.12	3.14	3.24	3.14
	s	0.79	0.61	0.70	0.82

The oneway ANOVA test and the Kruskal Wallis sum rank test results are as follows:

Tests of differences of the CEQ subscales by *ethnic group.*

	df	*F*	*p-value*	*df*	$KW\chi^2$	*p-value*
GT	3, 63.29	0.17	0.91	3	1.22	0.75
CG	3, 63.15	0.99	0.40	3	3.27	0.35
GS	3, 62.61	2.07	0.11	3	7.79	0.05
AA	3, 61.13	0.31	0.82	3	1.03	0.79
AW	3, 63.26	0.46	0.71	3	1.20	0.75

The results of the tests of differences show no evidence to suggest actual differences. This applies similarly to all five factors of the CEQ. These results imply that participants from the different ethnic background perceive their training experience similarly. This is true as far as the five factors of the CEQ are concerned.

8.5 Age Groups

In this section we look at the age of the respondents as a possible factor that is associated to the scales. Age is grouped into ranges which are similar to those in the evaluation form currently in used in INTAN.

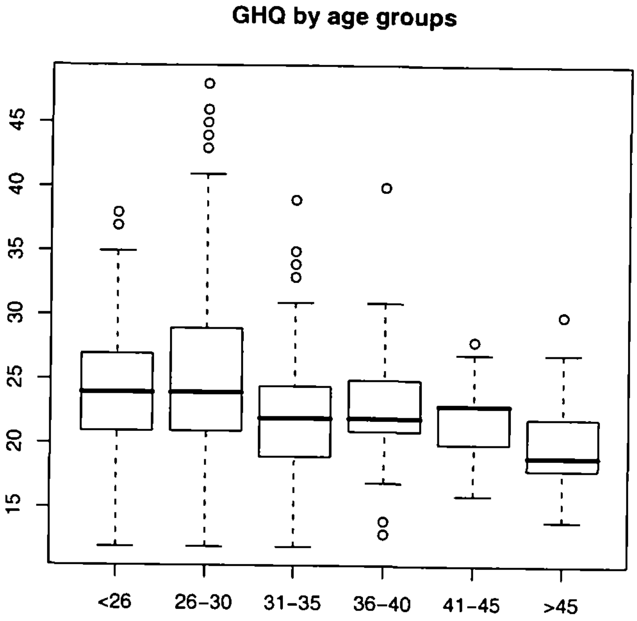


Figure 8.9: Boxplots of the GHQ by *age group.*

8.5.1 General Health Questionnaire

Figure 8.9 shows that the medians of the general health among the different age groups are not the same. The age group of > 45 years seems to have the lowest median, while medians for the age groups of < 26 years and 26 - 30 years seem to be about equally positioned. There also seems to be a decreasing trend of the GHQ score as the age increases. As lower score signifies better mental health, the trend implies that participants in the older groups perceive themselves as mentally healthier than those in the younger groups perceive theirs.

The interquartile ranges appear to decrease by age, suggesting decreasing variation in the score from the 'younger' participants to the 'older' participants. If this observed trend proves to be true, then it corresponds with the findings of a study among employees in a Japanese worksite [82]. The report by Shimizu also cites another similar finding in a different study among employees in Japanese companies.

Goldberg has reported in the manual [35] that there is no clear effect of age on the score of the GHQ. The finding which is reported in the manual is based on the GHQ-60, which is a much longer version than the GHQ-12 used in this study. Nevertheless, both versions of the questionnaire have been shown to be consistent with each other.

For this analysis, the number of samples, the means and the standard deviations of the GHQ by age groups are as follows:

Statistics of the GHQ by *age group.*

	<26	*26-30*	*31-35*	*36-40*	*41-45*	*>45*
n	185	357	128	30	31	26
$\overline{x}$	24.03	24.95	22.27	23.03	21.87	19.81
s	4.47	6.04	5.15	4.95	2.63	3.58

The Oneway ANOVA and Kruskal-Wallis tests produce the following results:

Test of differences of the GHQ by *age group.*

	Statistics	*p-value*
ANOVA	$F_{5,114.00} = 13.73$	1.73e-10
Kruskal Wallis	$W_5 = 49.44$	1.8e-09

At least two pairs of population means of the GHQ are different. Kruskal-Wallis rank sum test test emphasises the conclusions of ANOVA; scores of the GHQ among the population age groups are not equal.

Tukey HSD multiple comparison test shows that the significant differences are between the following age groups: (i) 31 - 35 and < 26, (ii) > 45 and < 26, (iii) 31 - 35 and 26 - 30, (iv) 41 - 45 and 26 - 30, and (v) > 45 and 26 - 30. These show that the significant differences are between two broader groups of age: the 'younger' groups of < 26 and 26-30 years, and the 'older' groups which consists of the other four age groups. The age of '30' seems to be the turning point where 'young' participants turns into 'old' ones and have different levels of general health.

8.5.2 Pretest Learning

Figure 8.10 shows that there is not that much variation in the scores of the LQ subscales among the age groups. In the case of the importance subscale, the distributions of the scores are close to the ceiling value. The sample sizes, the means

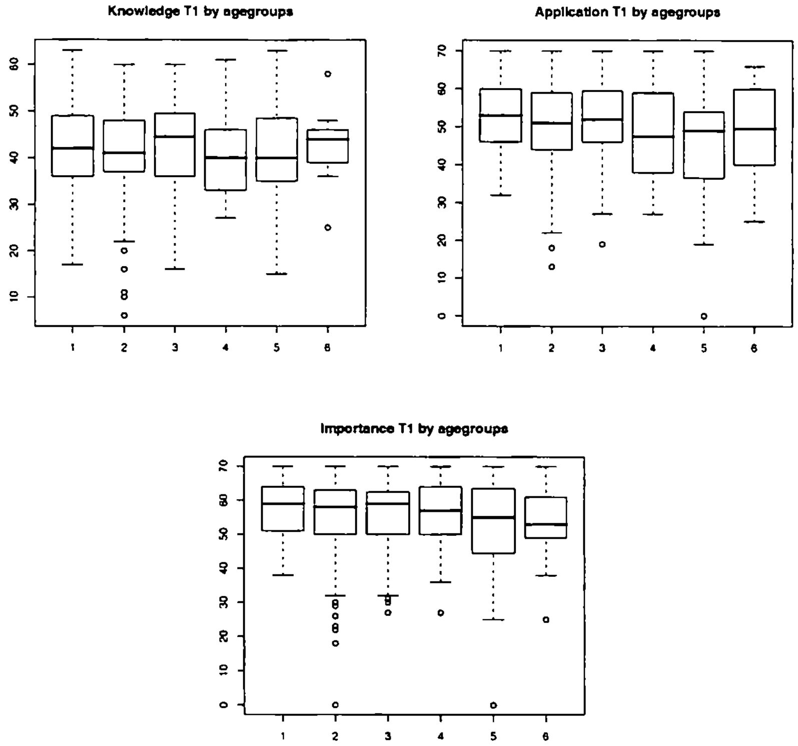


Figure 8.10: Pretest LQ subscales by *age group*.

and the standard deviations of the pretest LQ subscales by age groups are in the following summary statistics table.

Summary statistics of pretest LQ subscales by *age group*.

	Age groups	*<26*	*26-30*	*31-35*	*36-40*	*41-45*	*>45*
	n	185	357	128	30	31	26
Knowledge	$\overline{x}$	42.26	41.72	42.74	40.57	40.10	41.73
	s	8.70	8.43	8.90	8.24	12.59	7.57
Application	$\overline{x}$	52.49	51.21	51.66	48.87	45.00	48.73
	s	9.52	10.27	10.25	11.94	12.23	12.05
Importance	$\overline{x}$	57.58	56.01	56.23	55.33	54.10	53.50
	s	8.55	10.67	9.32	11.42	12.71	10.14

The ANOVA and the Kruskal-Wallis tests produce summarized outputs as in the following table:

Test of differences of pretest LQ subscales by age group.

	ANOVA			Kruskal		
	$df_{num,denom}$	F	p	df	χ^2	p
Knowledge	5, 106.5	0.60	0.70	5	3.94	0.56
Application	5, 103.99	1.98	0.09	5	9.68	0.08
Importance	5, 104.67	1.45	0.21	5	4.49	0.48

Both the parametric and non-parametric tests come to similar conclusions. In each of the three LQ subscales, the means of Learning factors from the different age groups do not differ significantly. In other words, participants from the different age groups do not indicate different level of pretest Learning.

8.5.3 Posttest Learning

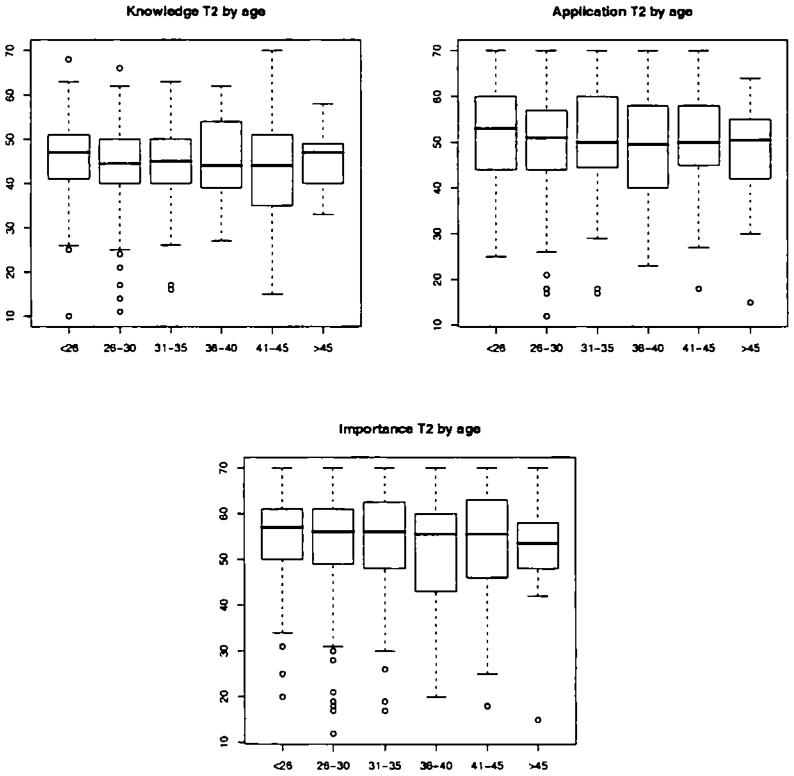


Figure 8.11: Posttest LQ subscales by *age group*.

Figure 8.11 suggest that the medians of the LQ subscales are similar among the age groups. There are some outliers on the low end of some of the groups. In the knowledge and importance subscales, the ceiling effect is quite obvious. The sample sizes, the means and the standard deviations are presented in the following summary statistics table. The results of the tests of difference are presented in a table following that.

Summary statistics of posttest LQ subscales by age groups.

	Age groups	*<26*	*26-30*	*31-35*	*36-40*	*41-45*	*>45*
	n	185	357	128	30	31	26
Knowledge	$\overline{x}$	46.28	44.40	44.86	45.33	42.48	46.08
	s	8.52	8.57	8.89	9.85	13.48	6.45
Application	$\overline{x}$	52.01	50.29	50.86	49.37	48.80	48.69
	s	9.73	10.70	11.07	11.52	12.17	10.45
Importance	$\overline{x}$	55.45	54.68	54.48	52.67	53.57	52.58
	s	9.86	10.88	11.10	12.59	13.57	10.37

Test of differences of posttest LQ subscales by *age group*.

	ANOVA			**Kruskal**		
	$df_{num,denom}$	F	p	df	χ^2	p
Knowledge	5, 103.08	1.49	0.20	5	6.55	0.26
Application	5, 104.96	1.16	0.33	5	4.46	0.48
Importance	5, 104.61	0.62	0.68	5	2.40	0.79

Just like the pretest data, the posttest Learning does not seem to be associated with age factor. There is no evidence to indicate that posttest Learning is different among the different age groups.

8.5.4 Course Experience Questionnaire

Figure 8.12 shows the distributions of the CEQ subscales by age groups. In some of the boxplots, the granularity effect as a result of the scales discrete nature is obvious. The sample sizes, the means and the standard deviations of each of the CEQ subscales by age groups are presented in the summary statistics table below.

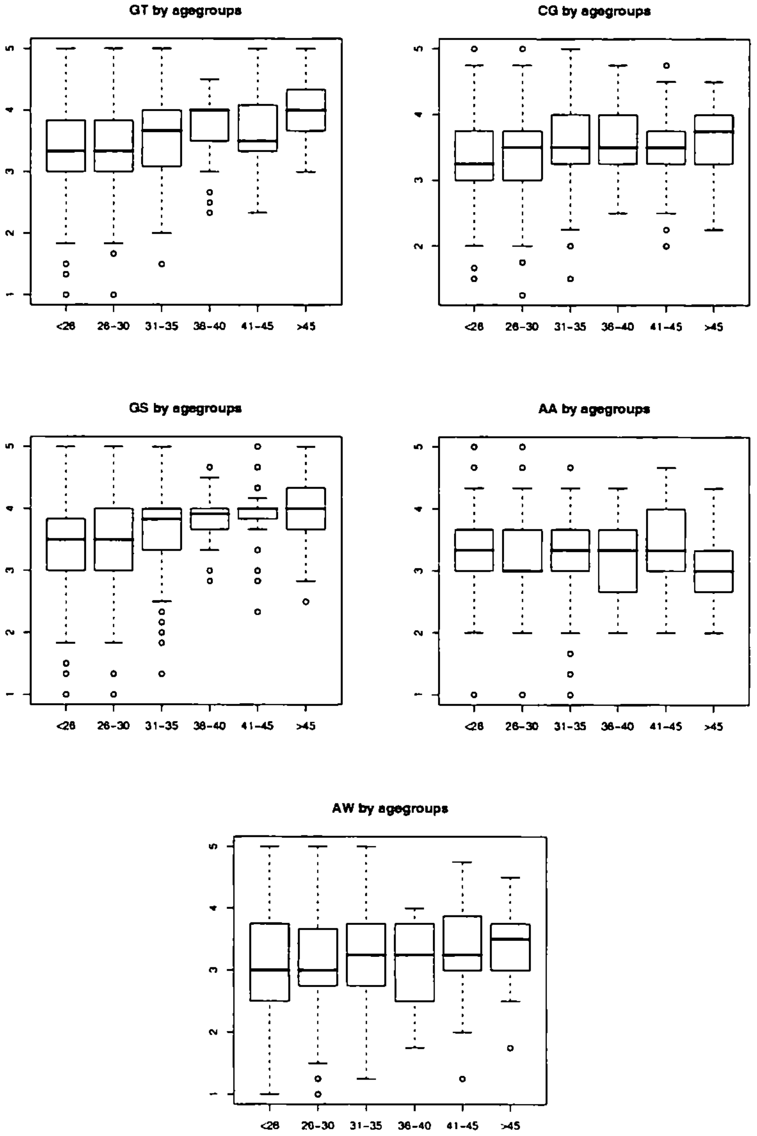


Figure 8.12: CEQ by *age group*.

Summary statistics of the CEQ subscales by *age group*.

Age groups		*<26*	*26-30*	*31-35*	*36-40*	*41-45*	*>45*
	n	185	357	128	30	31	26
GT	$\bar{x}$	3.33	3.39	3.53	3.73	3.70	3.98
	s	0.68	0.61	0.63	0.59	0.68	0.50
CG	$\bar{x}$	3.37	3.41	3.52	3.62	3.52	3.62
	s	0.57	0.52	0.57	0.54	0.72	0.61
GS	$\bar{x}$	3.42	3.48	3.69	3.88	3.90	3.98
	s	0.68	0.60	0.63	0.42	0.57	0.56
AA	$\bar{x}$	3.20	3.22	3.22	3.19	3.45	3.04
	s	0.62	0.58	0.66	0.60	0.74	0.54
AW	$\bar{x}$	3.10	3.08	3.22	3.02	3.31	3.46
	s	0.80	0.76	0.81	0.72	0.73	0.63

The ANOVA and the Kruskal-Wallis tests are carried out to test the hypotheses that the mean scores of the CEQ subscales are equal among the age groups. The results are presented in the following table.

Tests of differences of the CEQ subscales by *age group*.

	ANOVA			**Kruskal-Wallis**		
	$df_{num,denom}$	F	*p-value*	*df*	$KW\chi^2$	*p-value*
GT	5, 108.14	9.99	6.96e-08	5	40.9	6.69e-08
CG	5, 105.36	2.42	0.04	5	15.5	0.008
GS	5, 109.95	12.28	1.82e-09	5	54.5	1.69e-10
AA	5, 106.55	1.19	0.32	5	6.7	0.25
AW	5, 108.56	2.51	0.03	5	11.7	0.04

The results of both the parametric and the non-parametric tests suggest evidence of differences in the scores of two of the CEQ subscales among the age groups. The two subscales are the GT and the GS. There is also a weaker evidence for a significant difference in the CG subscale.

The GT subscale measures the perception of the participants about good teaching practices among the teaching staff or trainers, while the GS subscale measures the extent to which participants perceive the courses as building the necessary skills, namely *problem-solving, analytic skills, teamwork, confidence* and *communication skills* [96]. Participants from the different age groups have different perceptions on these two factors.

For the GT subscale, the significant differences are between the age groups of (i) 36-40 and < 26, (ii) 41-45 and < 26, (iii) > 45 and < 26, (iv) > 45 and 26-30, and (v) > 45 and 31-35. For the GS subscale, the age groups with significant differences are (i) 31-35 and < 26, (ii) 36-40 and < 26, (iii) 41-45 and < 26, (iv) > 45 and < 26, (v) 31-35 and 26-30, (vi) 36-40 and 26-30, (vii) 41-45 and 26-30, and (viii) > 45 and 26-30. Clearly, most of the differences involve the age groups of < 26 years and the 26-30 years old. This finding is similar to that in Section 8.5.1, where the age of 30 years seems to be the borderline. Here, it borders between two differing good teaching and generic skills scores.

It can be observed from Figure 8.12 that for the same two CEQ subscales, the first three age groups (< 26, 26-30, and 31-35 years) have wider dispersions that the other three age groups. This suggests that the differences in opinion in terms of the GT and the GS subscales are wider among the participants in these age groups, than the differences within the 'older' three age groups.

For the other three CEQ factors evidence of significant differences are not clear. There seem to be weak evidence of a significant difference for the CG subscale, but the probability value is large enough for the difference to be attributed to chance.

8.6 Centre

Examining the differences in the scores of the scales among the INTAN centres is a special interest of this study. Each centre of INTAN functions by running and managing training programmes of specific areas. Some of the centres even have specific participants as their target groups. It would be interesting to see whether the training programmes from the different centres bring about different impact in the participants with regards to the factors studied. It would also be of interest to the management of INTAN if participants from some centres get significantly different training experience from the participants of the other centres.

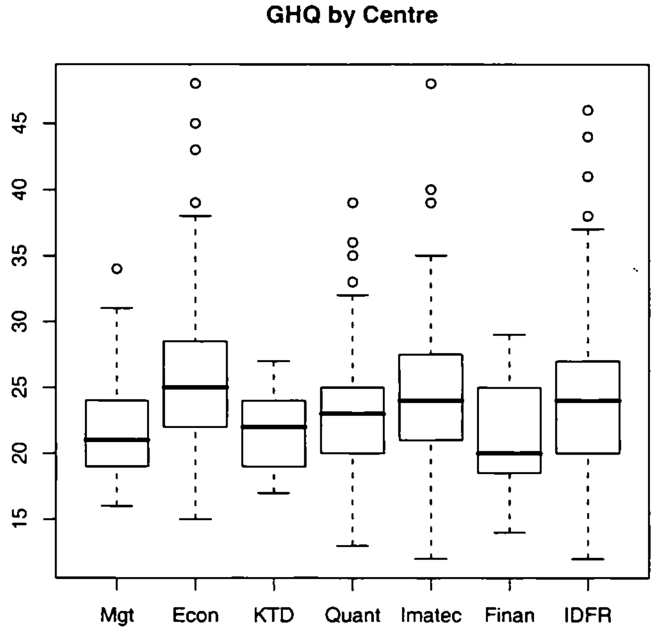


Figure 8.13: GHQ by *centre.*

8.6.1 General Health Questionnaire

Boxplots in Figure 8.13 suggest that there are differences in both the medians and the interquartile ranges of the GHQ scores among the different centres. Many of the distributions seem to be skewed to the right as well. The actual values of the means and the standard deviations of the GHQ by the centres are presented in the following table of summary statistics.

Summary statistics of the GHQ by *centre.*

	Management	*Economy*	*KTD*	*Quantitative*	*Imatec*	*Finance*	*IDFR*
n	26	155	8	172	155	28	216
$\overline{x}$	22.58	25.37	21.75	22.72	24.38	21.22	24.01
s	4.78	5.74	3.37	4.69	5.57	3.95	5.74

The results of the ANOVA and Kruskal-Wallis tests in the table below do imply that the means of the GHQ are not the same among participants from the different centres. However, the Tukey HSD test indicates that there are only two pairs of centres that have significant differences in the GHQ. The pairs are (i) the centre of Quantitative and the centre of Economy, and (ii) the centre of Finance and the centre of Economy.

Test of differences of the GHQ by *centre.*

	Statistics	*p-value*
ANOVA	$F_{6,72.47}$ = 6.09	3.34e-05
Kruskal Wallis	W_6 = 31.19	2.33e-05

Figure 8.13 shows that the centre of Economy has a higher GHQ score that do either the centre of Quantitative or the centre of Finance. The difference between the GHQ scores of the centre of Economy and the centre of KTD is also large, but the difference is not statistically significant, probably due to the small sample size from the centre of KTD.

The data shows that the participants from centre of Finance has the lowest mean, which implies that in general they have the highest level of mental health. The highest mean is from the centre of Economy, suggesting that in general participants from this centre are in the lowest state of mental health.

8.6.2 Pretest Learning

Looking at Figure 8.14, it can be inferred that the equality of the medians of the LQ subscales among the centres may not hold true. There could also be some variation in the interquartile ranges. The means and the standard deviations of the three pretest LQ subscales for each of the centres are as in the following table of summary statistics:

Summary statistics of pretest LQ subscales by *centre.*

	Centres	*Mgt*	*Econ*	*KTD*	*Quant*	*Imatec*	*Finan*	*IDFR*
	n	26	155	8	172	155	28	216
Knowledge	$\overline{x}$	39.38	40.09	39.75	41.60	41.98	40.57	44.04
	s	11.75	8.27	10.74	8.03	8.89	8.16	8.74
Application	$\overline{x}$	44.85	52.46	55.29	50.94	50.41	51.10	51.87
	s	12.51	10.36	10.08	10.16	10.49	8.89	10.02
Importance	$\overline{x}$	51.19	57.15	61.14	56.95	54.75	60.07	56.00
	s	11.64	9.43	7.49	9.88	10.75	8.91	9.88

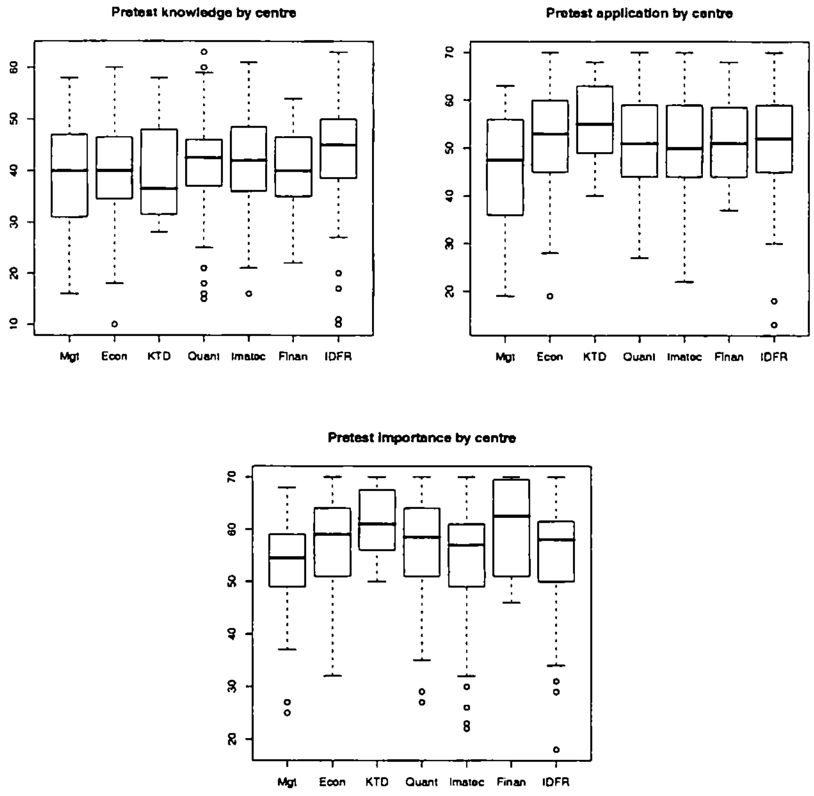


Figure 8.14: Pretest LQ subscales by *centre*.

The statistics from the t-test and Kruskal-Wallis tests are as follows:

Test of differences of pretest LQ subscales by centre

	ANOVA			Kruskal		
	$df_{num,denom}$	F	p	df	χ^2	p
Knowledge	6, 70.31	3.64	0.003	6	24.8	0.0004
Application	6, 64.16	1.90	0.09	6	11.88	0.06
Importance	6, 64.70	2.90	0.01	6	14.79	0.02

The results of the tests indicate a weak evidence for differences in knowledge. This suggests participants from at least one of the centres evaluated their level of knowledge differently from participants of the other centres. A post hoc test indicates that the significantly different mean of knowledge is between the centre of IDFR and the centre of Economy. The centre of IDFR has a mean score of 44.04, compared to the centre of Economy with 40.09, but the evidence to reject the hypothesis of equal means is not very strong. Other than that, there is no evidence of significant differences are observed for the application and importance subscales. These results are conflicting with the boxplots in Figure 8.14, which clearly suggest differences. It could be that because of the ceiling effect in the scores and the differences in the variances, ANOVA is not an effective test of differences.

8.6.3 Posttest Learning

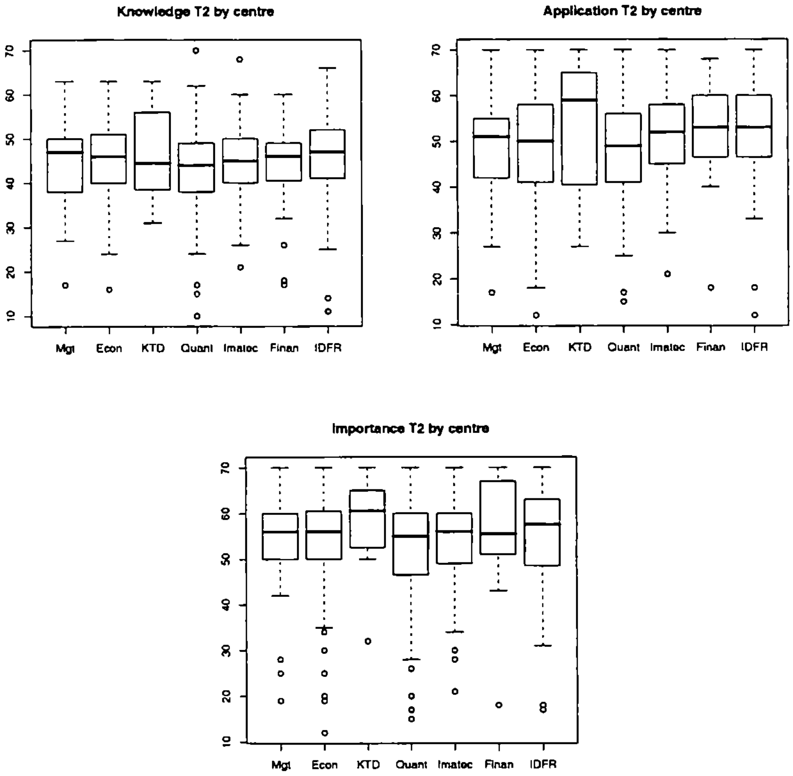


Figure 8.15: Posttest LQ subscales by *centre*.

Boxplots of the posttest LQ subscales in Figure 8.15 suggests not much variation in the medians, as well as in the interquartile ranges among the centres. A lot of outliers can also be seen, especially in the sores of importance, due to most scores being very high. The summary statistics of the scores are in the following table:

Summary statistics of posttest LQ subscales by *centre.*

	Centres	*Mgt*	*Econ*	*KTD*	*Quant*	*Imatec*	*Finan*	*IDFR*
	n	26	155	8	172	155	28	216
Knowledge	$\overline{x}$	43.77	45.05	46.50	43.28	44.71	43.75	46.61
	s	10.45	9.06	11.54	9.16	8.10	10.14	8.32
Application	$\overline{x}$	47.04	49.58	53.25	48.52	51.10	52.00	52.90
	s	12.38	11.20	16.18	10.85	9.80	10.04	9.76
Importance	$\overline{x}$	52.65	54.51	57.25	53.04	54.81	56.68	55.70
	s	12.46	11.11	12.07	11.51	10.28	11.36	10.04

Test of differences of posttest LQ subscales by *centre*

	ANOVA			**Kruskal**		
	$df_{num,denom}$	F	p	df	χ^2	p
Knowledge	6, 70.07	2.42	0.04	6	15.22	0.02
Application	6, 70.16	3.57	0.004	6	21.46	0.002
Importance	6, 70.41	1.22	0.31	6	6.28	0.39

The results of the tests show only one instance of an evidence for a significant difference, ie. of the application subscale. This suggests that the mean scores of the posttest application differ among the participants from the different centres, but the evidence is not strong. No evidence is indicated for the knowledge and importance subscales. The ceiling effect is more obvious with the posttest importance.

8.6.4 Course Experience Questionnaire

Figure 8.16 indicates variation in both the medians and the interquartile ranges among the seven centres. The variation in medians seems to be larger in the GT and AW subscales. For the GT subscale, the centre of Management and the centre of Finance appear to have high medians, while the centre of IDFR seems to have the lowest. The centre of IDFR also seems to have the lowest medians in CG and GS.

The medians of the AW are the most varied among the five subscales. Those of the centre of Economic and the centre of Quantitative are the lowest, while those of the centre of IMATEC and the centre of Finance are among the highest. The interquartile ranges also seem to vary a lot. The number of samples, the means

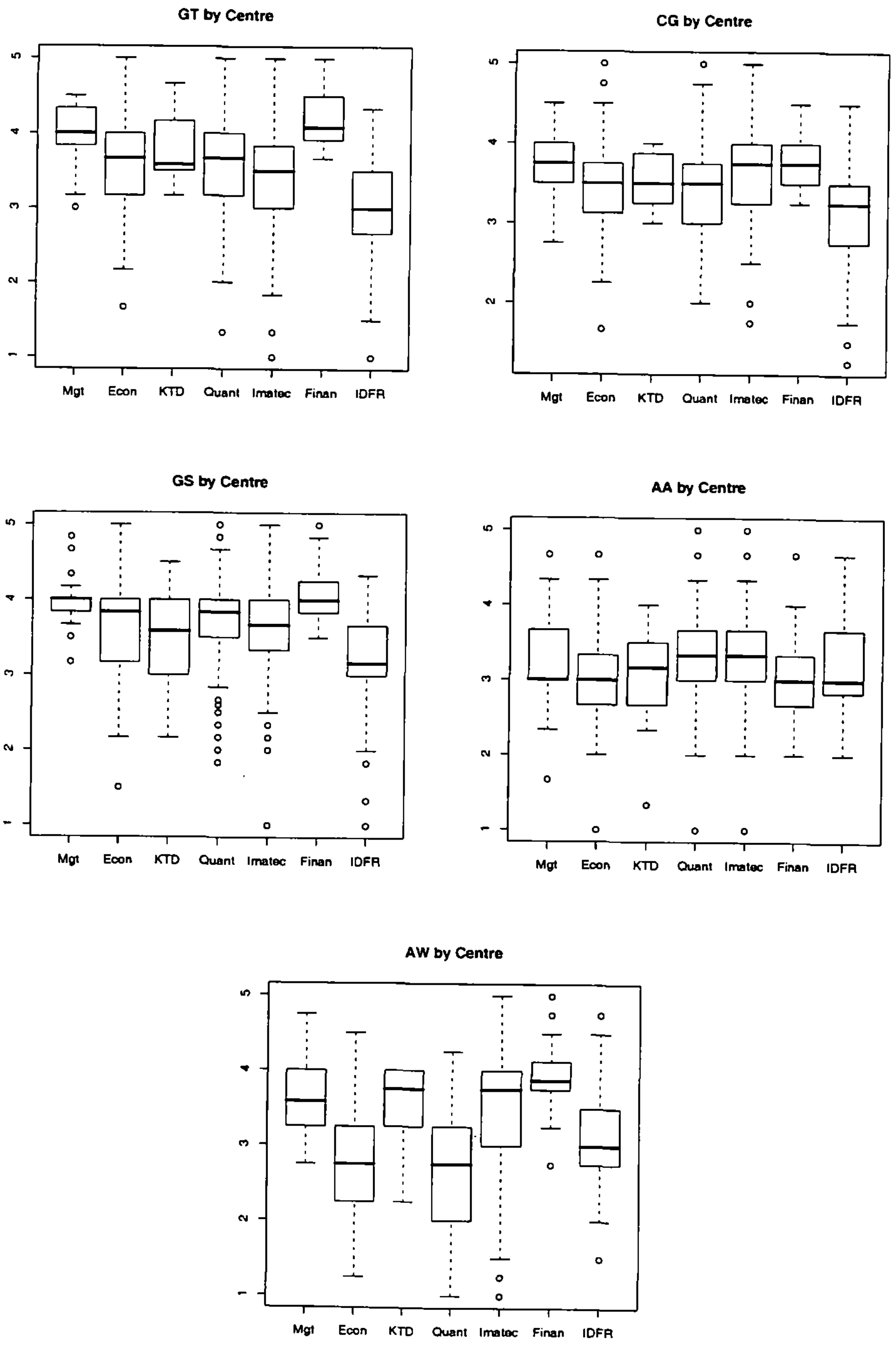


Figure 8.16: CEQ by *centre*.

and the standard deviations of the factors are as the following table of summary statistics:

Summary statistics of the CEQ subscales by *centre*.

centre		*Mgt*	*Econ*	*KTD*	*Quant*	*Imatec*	*Finan*	*IDFR*
	n	26	155	8	172	155	28	216
GT	$\bar{x}$	4.03	3.55	3.79	3.60	3.43	4.20	3.08
	s	0.40	0.60	0.50	0.60	0.60	0.39	0.59
CG	$\bar{x}$	3.68	3.48	3.53	3.47	3.62	3.81	3.17
	s	0.40	0.51	0.36	0.59	0.49	0.33	0.55
GS	$\bar{x}$	3.97	3.63	3.48	3.73	3.63	4.12	3.17
	s	0.33	0.60	0.78	0.59	0.57	0.39	0.61
AA	$\bar{x}$	3.27	3.10	3.00	3.34	3.27	3.02	3.18
	s	0.74	0.59	0.84	0.65	0.64	0.64	0.53
AW	$\bar{x}$	3.66	2.84	3.53	2.67	3.57	3.90	3.20
	s	0.52	0.71	0.60	0.74	0.78	0.51	0.57

The statistical tests of differences among the centres give the following results:

Tests of differences of the CEQ subscales by *centre*.

	ANOVA			**Kruskal-Wallis**		
	$df_{num,denom}$	F	*p-value*	*df*	$KW\chi^2$	*p-value*
GT	6, 72.73	40.14	2.2e-16	6	151.24	2.2e-16
CG	6, 73.26	18.58	5.31e-13	6	88.07	2.2e-16
GS	6, 72.54	30.86	2.2e-16	6	135.14	2.2e-16
AA	6, 70.06	2.63	0.02	6	20.07	0.003
AW	6, 71.82	37.80	2.2e-16	6	178.10	2.2e-16

Just as suggested by the boxplots in Figure 8.16, strong evidence for significant differences are found for all of the CEQ subscales, except the AA. A general conclusion that can be made from this is that participants attending training programmes under the different centres have significantly varied experience. They have different views regarding whether the trainers have good teaching practice, and they differ in whether they have clear understanding of what is expected of them during the training programmes. They also have different perceptions about whether the programmes are helping them build the necessary skills, and they differ in their views whether the amount of workload they have is appropriate. The only aspect in which they do not have differing views is the AA, which relates to whether the assessments of the training are appropriate.

Findings for the GT subscale imply that the teaching practice among the centres do differ. Looking at the values of the means, the centre of Finance has the highest score, followed by the centre of Management. The centre of IDFR has the lowest mean, which suggests that in the opinion of the respondents, trainers from this centre do not have such good teaching practices as those from the other centres.

The CG subscale measures the agreement of the participants that they are clear about what is expected of them during the programme. A high score implies that they are clear about the goals and the objectives of the programme. The results of this study indicate that the levels of agreement differ significantly among the different centres. Participants from the centre of Finance and the centre of Management score the highest means, while those from the centre of IDFR score the lowest.

The subscale of GS indicate their agreement about whether the training programmes are targetting the necessary skills. The results of the tests suggest there are significant differences in this levels of agreement among participants from the different centres. The centre of Finance and the centre of Management have high mean scores, while the centre of IDFR has the lowest.

For the scale of AA, a low score indicates that participants feel that they are being tested more for memory than understanding. In this study the result indicates that the difference is not significant. There is no evidence to conclude that participants from different centres give significantly different scores to this subscale.

The subscale of AW measures the perception about the amount of work they need to get through during the training programmes. A low score indicates that there is so much work that it impedes understanding. There is a strong evidence that participants from different centres view this differently. The centre of Finance has the highest mean, followed by the centre of Management. The centre of Quantitative has the lowest mean.

Figure 8.17 shows the mean scores of each of the centres for all CEQ subscales. The centres are each represented by a line of different colours. Higher scores indicate more positive reaction from the study participants with regards to the scope of the

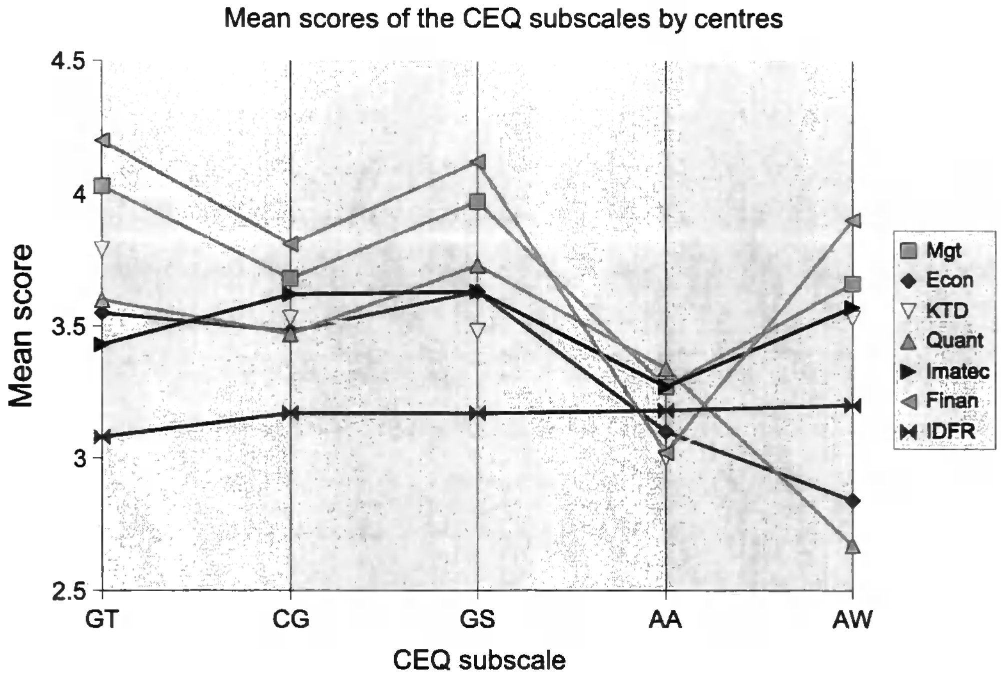


Figure 8.17: Mean scores of CEQ subscales by *centre*.

CEQ subscales. The position of the mean score of each centre in relation to the other centres for each subscale is clearly presented. The mean scores of four of the CEQ subscales, namely the GT, CG, GS, and AW, are widely varied, but the mean scores of the AA is not. This suggests that generally, participants from all centres feel differently about the four aspects of training, but they have similar feeling about the assessment aspect.

The centre of Finance seems to be the 'best' overall performer. It has the highest mean scores in four of the CEQ subscales, namely the GT, CG, GS, and AW. However, it scores among the lowest in the AA subscale. These indicate that participants from this centre are generally happy about the good teaching practice, the clear directions and expectations, the necessary skills targeted by the programmes, and the amount of work they have to go through, but they are not very happy about the training assessment. Of course, their feeling about the assessment is not much different from those of the other centres.

The second best performer is the centre of Economy. It has the second highest mean scores in the same four subscales. The centre's mean score for the AA is among the highest. These results indicate that the participants from this centre are also generally happy, except for the way they are assessed.

Another observation worth noting regards the centre of Quantitative. Its mean scores in the subscales of GT, CG, and GS seem typical, but it has the highest mean score for the AA subscale, and the lowest mean score for the AW subscale. The highest score for the AA implies that the participants from this centre is the happiest about how they are assessed. On the other hand the lowest mean score for the AW implies that the participants are least satisfied with the amount of work they have to do during the training programme.

The centre of IDFR shows a peculiar pattern of mean scores compared to the other centres. There is not much variation in the scores, and all mean scores are just above 3. In the scale of 1-5, a 3 implies several possibilities. First is that the respondents just 'couldn't be bothered' about the evaluation. Secondly, the respondents chose to be neutral on the items being evaluated, where they neither agree nor disagree, or they simply do not know.

In terms of the subscales, there is not much variation in the mean scores of the AA, but generally the scores are all low. It could be implying that the participants in general are not very clear of what is expected of them during the training. The situation seems worst with the centres of Finance and KTD, but not as bad with the centres of Quantitative, Economy, Management and IMATEC.

8.7 Service Sector

Federal, *state*, and *local* are the three levels of this factor. Most participants will be from the federal level, which consists of all the Ministries and Government Departments. State refers to the administration at state level, and local refers to the Local Authorities and Local Councils, which are often dubbed as the Local Government.

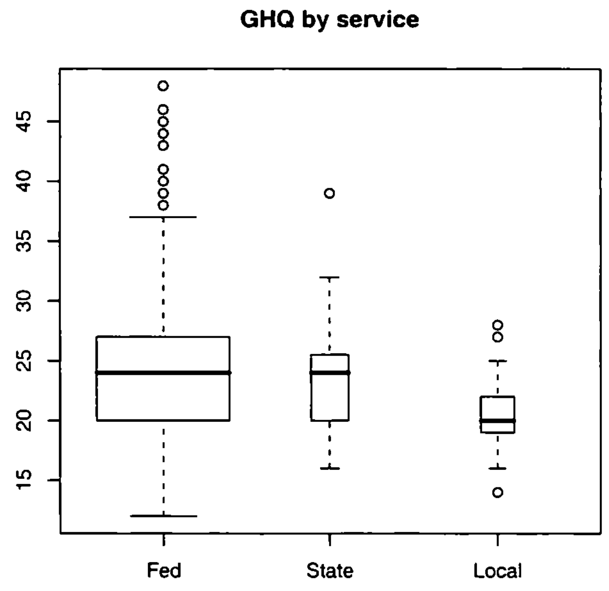


Figure 8.18: GHQ by *service sector*, variable width.

8.7.1 General Health Questionnaire

The number of samples are 664 (Federal), 52 (State) and 41 (Local), and the differences are reflected in the different widths of the boxplots in Figure 8.18. The figure shows that the GHQ score of the local is slightly lower than those of the other two service sectors. The means of the GHQ are 24.13 for federal, 23.56 for state, and 20.61 for local. Standard deviation values are 5.57, 4.66, and 3.15 for federal, state, and local respectively. Statistical tests for the hypothesis of no differences among the services sectors gave the following results:

Test of differences of the GHQ by *service sector.*

	Statistics		*p-value*
ANOVA	$F_{2,77.05}$	$= 21.22$	4.55e-08
Kruskal Wallis	W_2	$= 20.30$	3.91e-05

The results indicate evidence of significant differences among the service sectors in terms of the GHQ. A posthoc test confirms that the significant differences are between the *local* and the other two levels. The conclusion is that the participants working in the Local Governments have significantly low GHQ score compared to the other two groups. Since a low score of the GHQ means better health, this is a suggestion that participants working in the local government or local authorities are more healthy than their colleagues working in the other two sectors.

8.7.2 Pretest Learning

Figure 8.19 shows not much variation in the medians of the pretest LQ subscales among the service sectors. The interquartile ranges do not seem to vary very much either. The values of the mean and the standard deviation of the pretest LQ subscales by service sectors are presented in the table of summary statistics below:

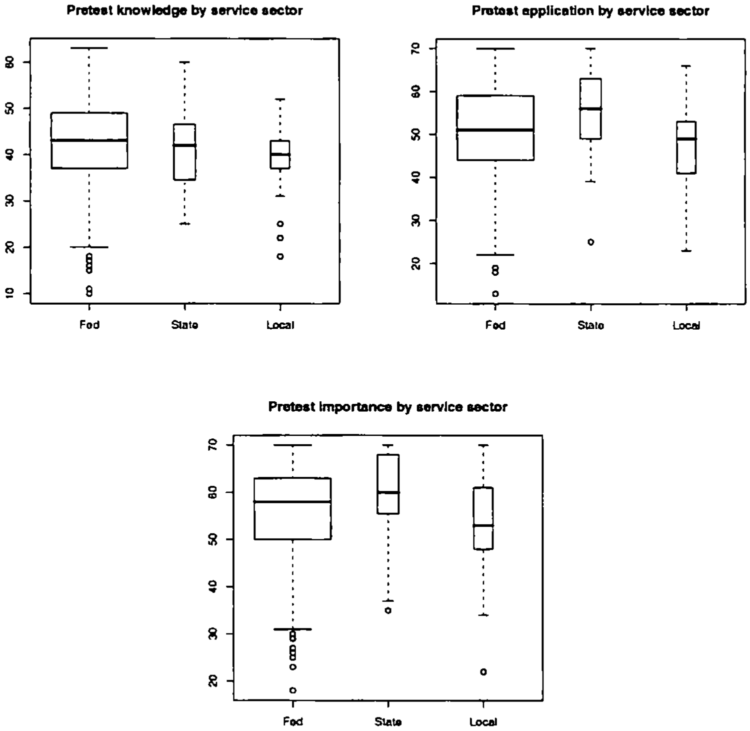


Figure 8.19: Pretest LQ subscales by *service sector*, variable width.

Summary statistics of pretest LQ subscales by *service sector*.

Service sector		*Federal*	*State*	*Local*
	n	664	52	41
Knowledge	$\overline{x}$	42.17	41.75	38.54
	s	8.82	8.41	7.28
Application	$\overline{x}$	51.15	55.16	47.83
	s	10.38	9.54	9.48
Importance	$\overline{x}$	56.19	59.45	53.10
	s	10.04	9.30	10.70

The means of the three LQ subscales are quite different among the three service types. The local service sector shows the lowest means in all three LQ subscales. The results of the statistical tests are as follows:

Test of differences of pretest LQ subscales by *service sector*.

	ANOVA			Kruskal		
	$df_{num,denom}$	F	p	df	χ^2	p
Knowledge	2, 69.87	4.63	0.013	2	7.94	0.02
Application	2, 68.47	6.91	0.002	2	11.84	0.003
Importance	2, 67.19	4.76	0.012	2	8.9	0.012

The results show that there is only a weak evidence for a significant difference in the application subscale. The hypothesis of no difference is clearly not rejected for the other two LQ subscales. It can be concluded that generally, participants from the different service sectors do not show much difference in their levels of knowledge, in their perception towards the application of the subjects, or in their perception towards the importance of learning the subjects.

8.7.3 Posttest Learning

Figure 8.20 indicates that there is not much variation in both the medians and the interquartile ranges of the posttest LQ subscales, among the service sectors. In all three subscales, the median of the state sector is highest, followed by those of the federal sector and the local sector. The interquartile range of the federal sector appears to be the largest in all three subscales, with outliers at the lower ends.

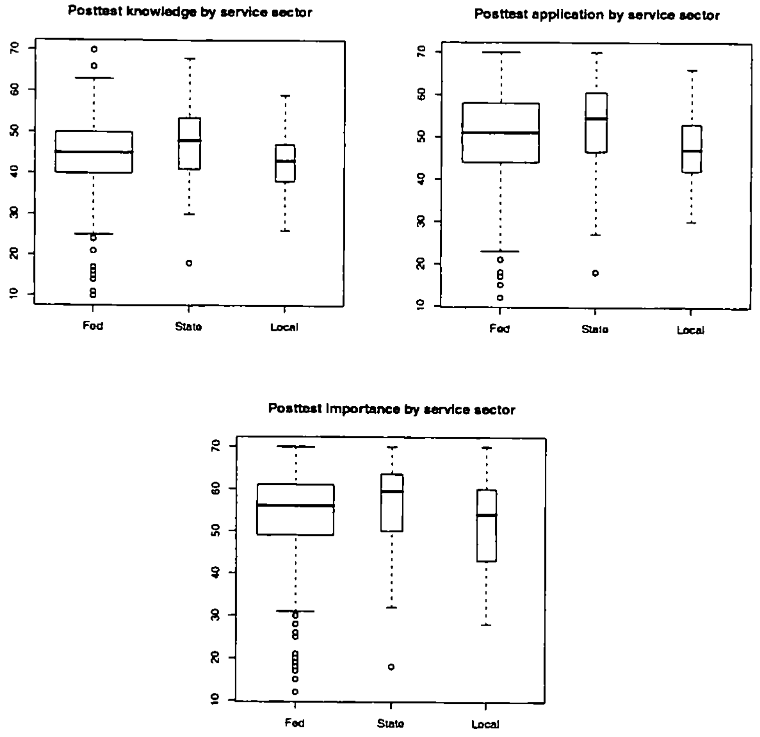


Figure 8.20: Posttest LQ subscales by *service sector*, variable width.

Summary statistics of posttest LQ subscales by *service sector.*

Service sector		*Federal*	*State*	*Local*
	n	664	52	41
Knowledge	$\overline{x}$	44.97	47.12	42.83
	s	8.88	9.23	7.11
Application	$\overline{x}$	50.69	52.92	47.90
	s	10.60	11.36	9.02
Importance	$\overline{x}$	54.59	57.23	52.05
	s	10.80	10.94	10.80

The statistical tests of differences done on the posttest LQ subscales produced the following results:

Test of differences of posttest LQ subscales by *service sector*.

	ANOVA			Kruskal		
	$df_{num,denom}$	F	p	df	χ^2	p
Knowledge	2, 69.63	3.22	0.05	2	7.35	0.03
Application	2, 68.75	2.92	0.06	2	6.96	0.03
Importance	2, 67.62	2.62	0.08	2	7.0	0.03

The results of the tests of differences show no evidence of significant differences in any of the posttest LQ subscales among the three service sectors. The average scores of the posttest knowledge, application, and importance are equal among the three service sectors.

8.7.4 Course Experience Questionnaire

Figure 8.21 suggests little variation in the medians of the CEQ subscales among the service sectors. The local service sector seems to have the highest scores in the GT, CG, GS and AW subscales. The federal sector appears to have the widest interquartile ranges in all subscales. The values of the means and the standard deviations are in the following table of summary statistics:

Summary statistics of the CEQ subscales by *service sector*.

Service sector		*Federal*	*State*	*Local*
	n	664	52	41
GT	$\bar{x}$	3.42	3.48	3.90
	s	0.65	0.55	0.54
CG	$\bar{x}$	3.42	3.50	3.68
	s	0.57	0.48	0.39
GS	$\bar{x}$	3.51	3.69	3.94
	s	0.64	0.54	0.48
AA	$\bar{x}$	3.21	3.24	3.23
	s	0.63	0.53	0.51
AW	$\bar{x}$	3.09	3.26	3.51
	s	0.78	0.78	0.66

The results of the oneway ANOVA and Kruskal-Wallis tests are in the following table.

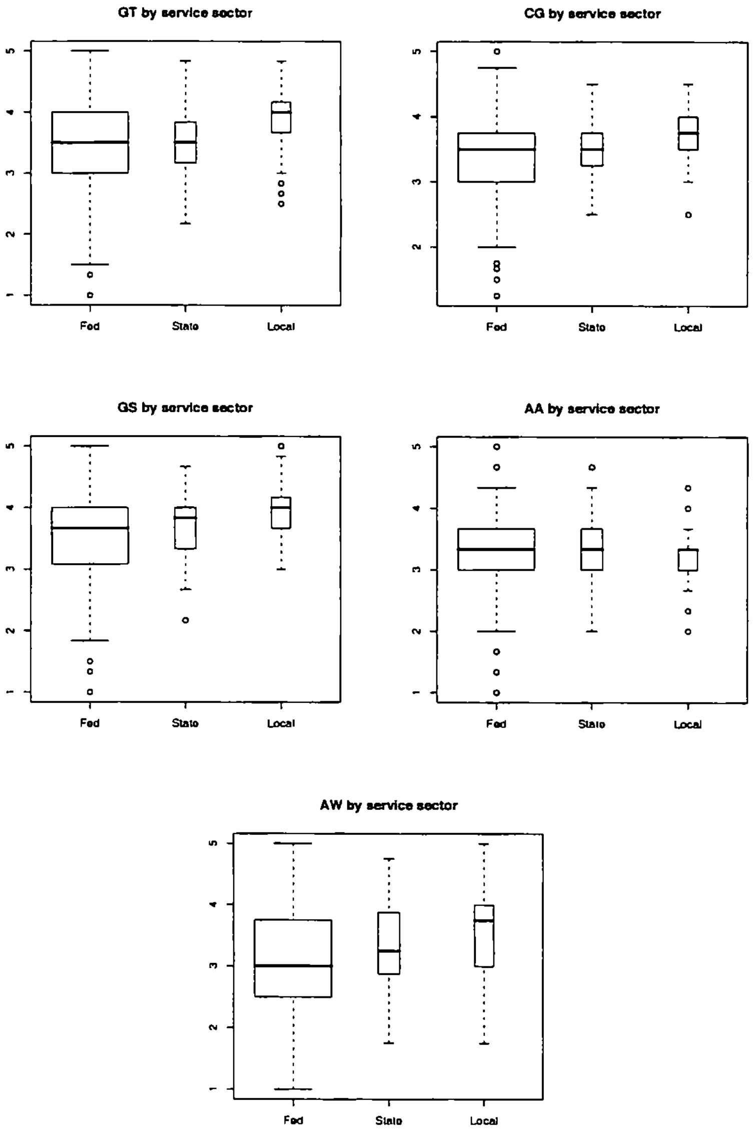


Figure 8.21: CEQ subscales by *service sector*, variable width.

Tests of differences of the CEQ subscales by *service sector*.

	ANOVA			Kruskal-Wallis		
	$df_{num,denom}$	F	*p-value*	*df*	$KW\chi^2$	*p-value*
GT	2, 70.84	14.89	3.99e-06	2	23.5	7.9e-06
CG	2, 73.28	8.43	0.0005	2	11.61	0.003
GS	2, 72.23	15.56	2.4e-06	2	18.24	0.0001
AA	2, 71.16	0.05	0.95	2	0.07	0.97
AW	2, 69.12	8.25	0.0006	2	13.83	0.001

The results indicate strong evidence of significant differences in the scores of the GT and GS subscales, and slightly less evidence in the scores of the CG and AW subscales. There is no evidence likewise in the score of the AA subscale. Participants from the different service sectors differ in their views with regards to the four CEQ subscales, ie. not including the AA.

The Tukey HSD post-hoc test shows that for the GT subscale, significant differences are observed between the sector pairs of Local - Federal and Local - State. For CG, GS and AW, the only significant difference is between Local - Federal. Participants from the local governments have views about the good teaching practice, which is different from those of the participants from the Federal and State governments. They also have views about the clear goals, generic skills, and appropriate workload factors which are different from those of the participants from the Federal government.

8.8 Service Group

The majority of participants who attend training programmes at INTAN are from the two main groups of service, namely the **Management and Professional** staff (Prof) and the **Supporting** staff (Supp). In INTAN, there are some training programmes that cater for these groups together, where both groups can attend at the same time. Other programmes cater specifically for either group, in which case only those from the appropriate group may apply.

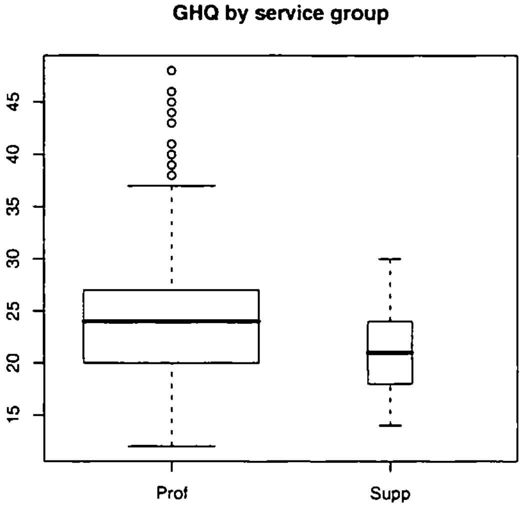


Figure 8.22: GHQ by *service group*, variable width.

8.8.1 General Health Questionnaire

Figure 8.22 shows a small difference in the medians between the two groups. The professional staff has a lot of outliers on the upper side of the distribution, which are the high scores that represent bad mental health condition. The interquartile ranges of the two groups also seem to differ.

The means of the GHQ are 24.09 for the professional and 21.26 for the supporting staff. The respective standard deviations are 5.52 and 3.61. The results of the t-test and Wilcoxon rank sum test tests are as follows:

Tests of differences of GHQ by *service group*.

Welch t-test				Mann-Whitney	
df	*t*	*p-value*	95%*CI*	*W*	*p-value*
59.67	4.95	6.32e-06	(1.69, 3.97)	21587	0.0002

The results of the tests show evidences of a significant difference in the mean GHQ score between the two service groups. Participants from the supporting staff indicate a slightly lower score, implying that they are better off than the professional staff, in terms of mental health [18]. However, the difference of the means is just 2.83.

In a scale of between 12 to 48, a difference of this size might not be practically significant.

8.8.2 Pretest Learning

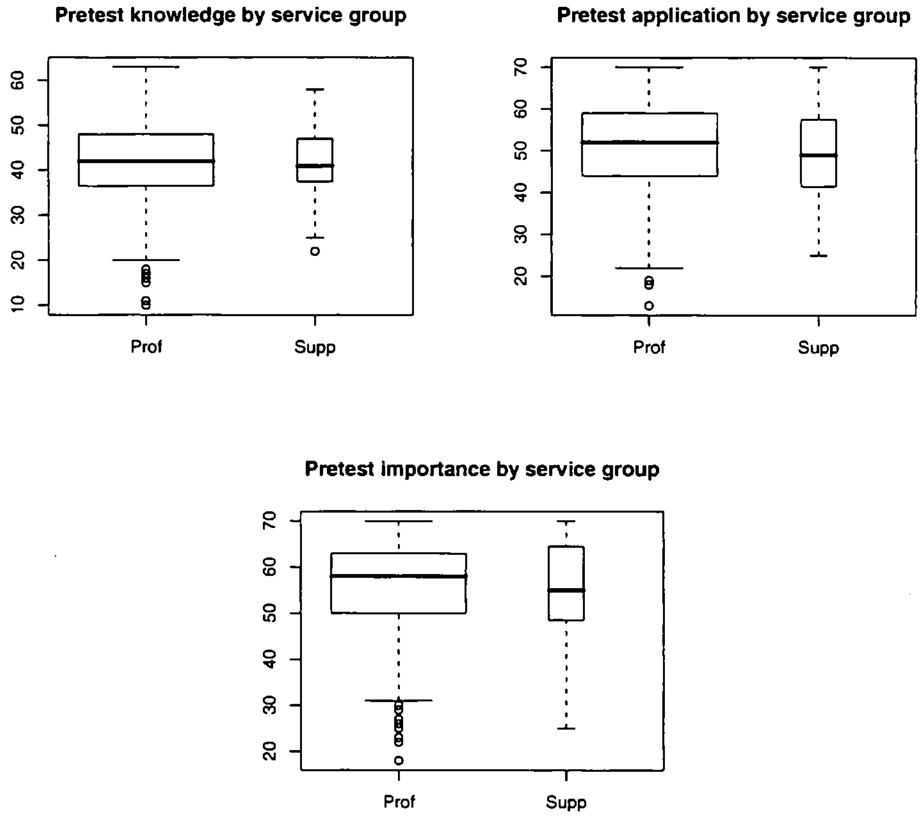


Figure 8.23: Pretest LQ subscales by *service group*, variable width.

Figure 8.23 suggests that the medians of the LQ subscales do not differ much between the professional and the supporting group. There are outliers at the low end of the scales of the professional group. The values of the means and the standard deviations of the three LQ subscales by the service groups are as presented in the following table of summary statistics:

Summary statistics of pretest LQ subscales by *service group*.

Service sector		*Professional*	*Supporting*
	n	708	47
Knowledge	$\overline{x}$	41.98	41.70
	s	8.79	8.34
Application	$\overline{x}$	51.32	48.62
	s	10.31	10.35
Importance	$\overline{x}$	56.30	54.89
	s	9.94	11.43

The results of the test of differences are as follows:

Test of differences of pretest LQ subscales by *service group*.

Statistics	*Knowledge*	*Application*	*Importance*
df	52.88	52.25	50.74
t	0.22	1.78	0.83
p-value	0.83	0.08	0.41
95%CI	(-2.25, 2.81)	(-0.35, 5.9)	(-2.02, 4.84)
W	16951.5	19410	17925
p-value	0.82	0.05	0.35

The results shows no evidence of significant differences in the scores of the pretest LQ subscales between the two groups. Participants from the professional group and the supporting group have similar mean scores of the knowledge, application, and importance. In other words, there is no indication that the pretest Learning of both service groups are not similar.

8.8.3 Posttest Learning

In Figure 8.24 we can see again the existence of outliers in the scores of the professional staff. Most of the outliers are on the low end of the scale, representing good mental health scores. The medians of the two groups do not seem to differ, but the interquartile ranges of application and importance obviously differ. The values of the means and the standard deviations are presented in the following table of summary statistics:

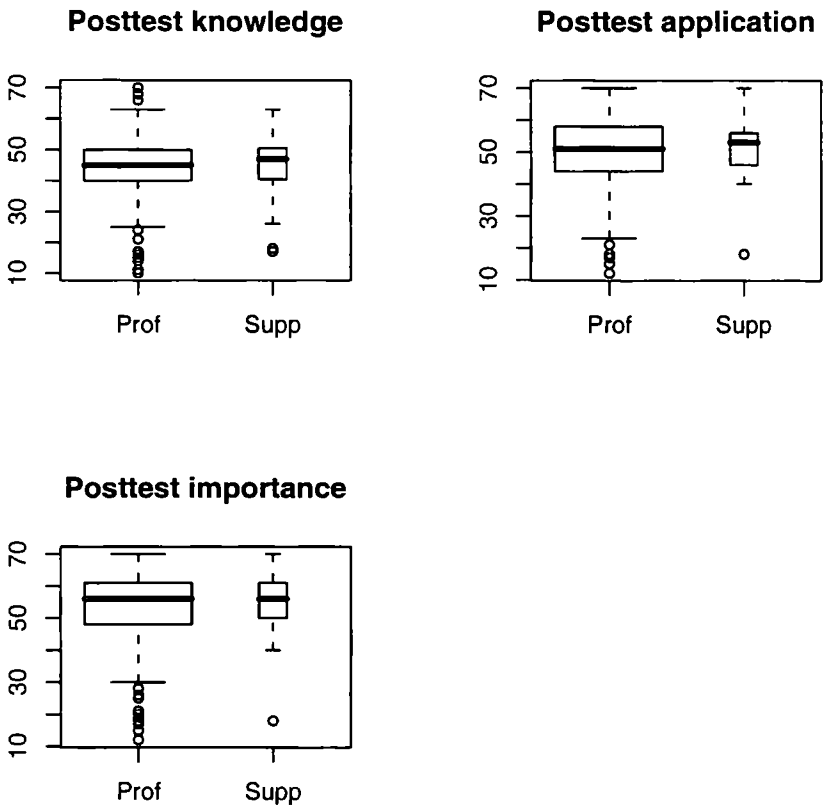


Figure 8.24: Posttest LQ by *service group*, variable width.

Summary statistics of posttest LQ subscales by *service group*.

Service sector		*Professional*	*Supporting*
	n	708	47
Knowledge	$\overline{x}$	44.92	45.32
	s	8.80	9.52
Application	$\overline{x}$	50.53	51.91
	s	10.67	9.44
Importance	$\overline{x}$	54.57	55.47
	s	10.86	10.17

The t-test and the Wilcoxon tests produced results as in the following table:

Test of differences of posttest LQ subscales by *service group*.

Statistics	*Knowledge*	*Application*	*Importance*
df	51.39	54.16	53.22
t	-0.28	-0.97	-0.58
p-value	0.78	0.34	0.56
95%CI	(-3.27, 2.46)	(-4.26, 1.45)	(-0.58, 2.19)
W	15433.5	15107	16110
p-value	0.44	0.32	0.74

The results of the test of differences are similar to those of the pretest LQ subscales; no evidence is found to support differences in the posttest LQ subscales between the two service groups. The posttest Learning of the professional staff and the supporting staff groups are similar.

8.8.4 Course Experience Questionnaire

Figure 8.25 shows differences in both the medians and the interquartile ranges between the service groups for all CEQ subscales. For the GT, CG, GS and AW subscales, the professional group seem to have higher medians, while for the AA subscale the supporting staff group scores higher. The interquartile ranges of the professional staff are wider in all cases, with outliers in all subscales except the AW.

The values of the means and the standard deviations of the CEQ subscales by the service groups are as in the following table of summary statistics. In the table following that, the results of the tests of differences are presented.

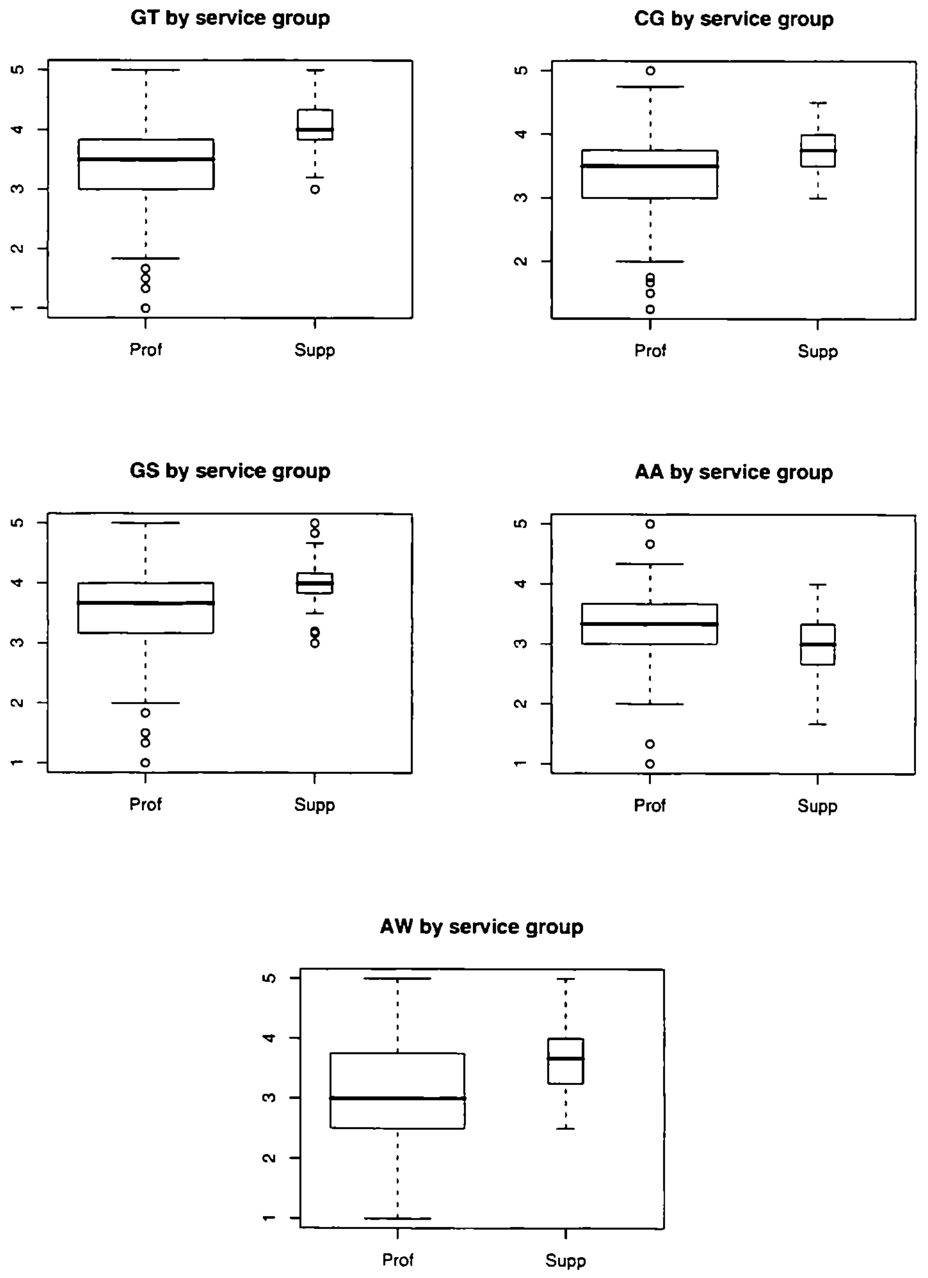


Figure 8.25: CEQ by *service group*, variable width.

Summary statistics of the CEQ subscales by *service group.*

	Sex	*Professional*	*Supporting*
	n	424	334
GT	$\overline{x}$	3.41	4.01
	s	0.63	0.46
CG	$\overline{x}$	3.42	3.72
	s	0.56	0.33
GS	$\overline{x}$	3.53	3.99
	s	0.63	0.45
AA	$\overline{x}$	3.23	2.91
	s	0.61	0.57
AW	$\overline{x}$	3.09	3.60
	s	0.78	0.58

Tests of differences of the CEQ subscales by *service group.*

	df	*t*	*p-value*	95% CI	*W*	*p-value*
GT	58.18	-8.40	1.26e-11	(-0.74, -0.46)	7316.5	1.01e-10
CG	65.38	-5.69	3.17e-07	(-0.40, -0.19)	10987	8.01e-05
GS	58.83	-6.68	9.31e-09	(-0.60, -0.33)	9489.5	6.85e-07
AA	53.32	3.70	0.0005	(0.15, 0.49)	21599.5	0.0005
AW	57.68	-5.76	3.49e-07	(-0.69, -0.33)	10204.5	7.96e-06

The results of these tests indicate clear evidence for significant differences in the scores of all five CEQ subscales. This means that generally, the professional and the supporting staff have different levels of experience while attending training programmes at INTAN.

Looking at the mean values, the supporting staff gave higher scores for four CEQ subscales, namely the GT, CG, GS, and AW. The only subscale for which they gave a lower score is the AA. Generally, this shows that training participants from the supporting staff on the average have more positive reaction towards the training programmes at INTAN than do the professional and management group. The only aspect on which they feel more negative compared to the professional group is regarding the assessment, where they feel that they are being tested more for their memory instead of their understanding [77].

8.9 Experience

The factor of experience refers to the number of years the respondents have been working in the public sector. The years are in ranges, and these ranges are similar to the ranges used in the current INTAN evaluation questionnaire.

8.9.1 General Health Questionnaire

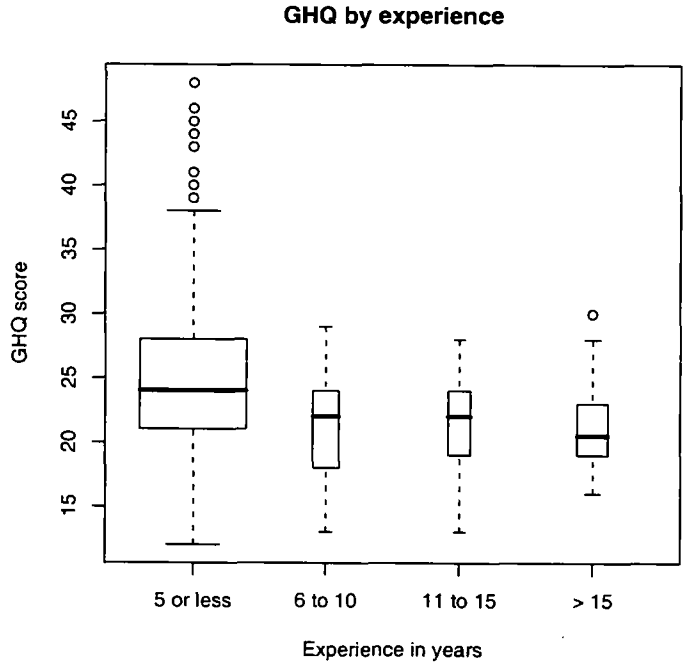


Figure 8.26: GHQ by *experience*, variable width.

Figure 8.26 suggests some differences in the medians among the experience levels, especially between that of the 5 years or less group, and those of the other experience levels. There are a few outliers at the upper end of the 5 years or less group of respondents. The mean and the standard deviation of each of the experience levels is as follows:

Means of GHQ by *experience*.

Experience	*5 yrs or less*	*6 to 10 yrs*	*11 to 15 yrs*	*> 15 yrs*
n	638	39	28	55
$\overline{x}$	24.41	21.33	21.18	21.11
s	5.60	3.92	4.26	3.29

The tests of differences on the GHQ by *experience* produced the following results:

Test of differences of the GHQ by *experience*.

	Statistics	*p-value*
ANOVA	$F_{3,73.1}$ = 21.41	4.69e-10
Kruskal Wallis	W_3 = 36.52	5.80e-08

The results of the tests show strong evidence of significant differences in the GHQ scores among the different experience levels. Looking at the mean values and the boxplots (Figure 8.26), participants with five years or less working experience have the highest mean score. This implies that this group has comparatively worse 'general mental health'. The more experienced the participants are, the more healthy they appear to be in terms of their general health. Figure 8.26 also indicates that the group with five years or less experience has a large interquartile range, while the group with the most experience, ie. those with over fifteen years experience, seems to have the smallest variability. Staff who have been in the job for many years are generally older than those who have only been working for a few years. Thus these findings would be expected if "more experience" means more stability, in terms of their job, as well as their family lives.

A post hoc test shows that the significant differences in the GHQ are between the experience groups of (i) 6 to 10 - 5 or less, (ii) 11 to 15 - 5 or less, and (iii) > 15 - 5 or less. Clearly, the group with experience of 5 years or less has significantly different general health score from any other experience groups.

8.9.2 Pretest Learning

Figure 8.27 shows that there is not much difference in the medians among the experience levels. However, the group with more than 15 years experience appear

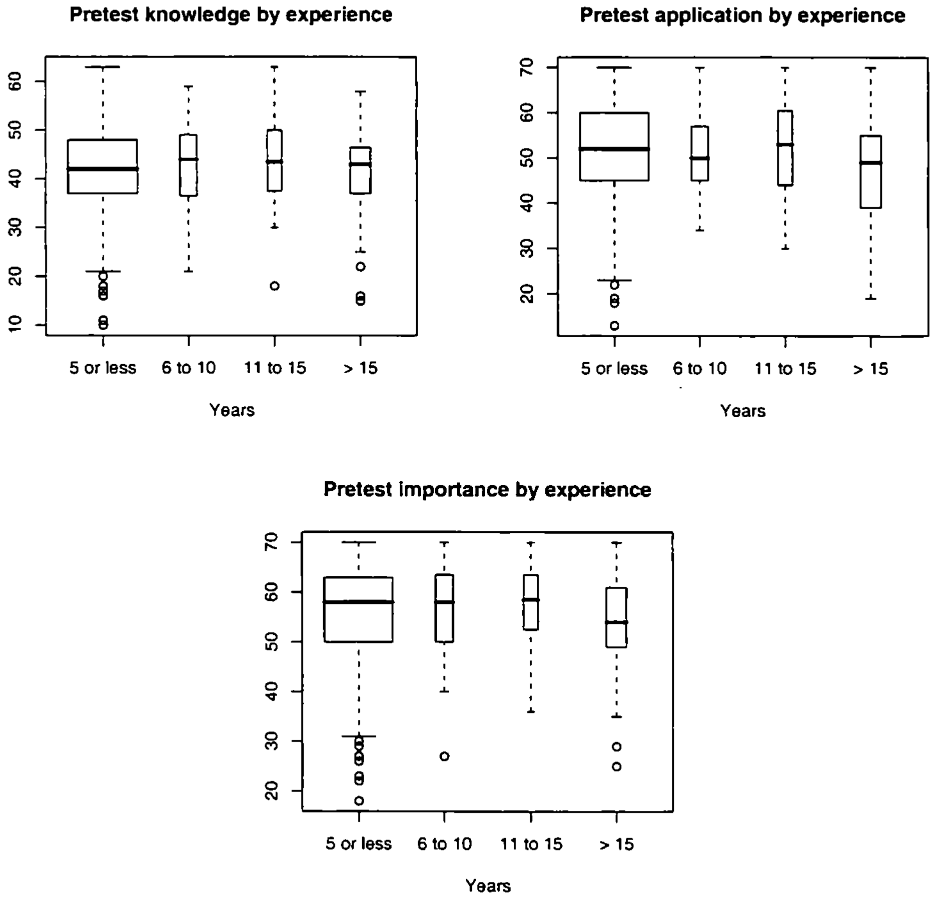


Figure 8.27: Pretest LQ subscales by *experience*, variable width.

to have the lowests scores in all three LQ subscales. The following table shows the summary statistics of the three subscales by experience.

Summary statistics of pretest LQ subscales by *experience.*

Experience (years)		*5 or less*	*6 to 10*	*11 to 15*	*> 15*
	n	638	39	28	55
Knowledge	$\bar{x}$	41.97	42.21	43.07	40.76
	s	8.59	8.76	10.23	9.69
Application	$\bar{x}$	51.58	51.21	51.43	47.15
	s	10.16	9.12	10.83	12.21
Importance	$\bar{x}$	56.39	56.67	57.43	53.37
	s	10.02	9.87	8.78	11.23

Statistical tests of differences give the following results:

Test of differences of pretest LQ subscales by experience

	ANOVA			**Kruskal**		
	$df_{num,denom}$	F	p	df	χ^2	p
Knowledge	3, 66.47	0.39	0.76	3	0.58	0.90
Application	3, 66.92	2.20	0.10	3	6.95	0.07
Importance	3, 67.59	1.38	0.26	3	4.02	0.26

The results of the statistical tests indicate no evidence of significant differences in the pretest LQ scores among the groups with different experience levels. Participants with different ranges of experience perceive their pretest knowledge to be equal. Similarly, they perceive the usage of the subjects to be equal (application), as well as how they feel about the importance of learning the subjects (importance).

8.9.3 Posttest Learning

There is not much variation in the medians of the posttest LQ subscales among the different experience levels as shown in Figure 8.28. The values of the means and the standard deviations of the posttest LQ subscales by experience are in the following table of summary statistics. The results of the tests of differences are in the table following that.

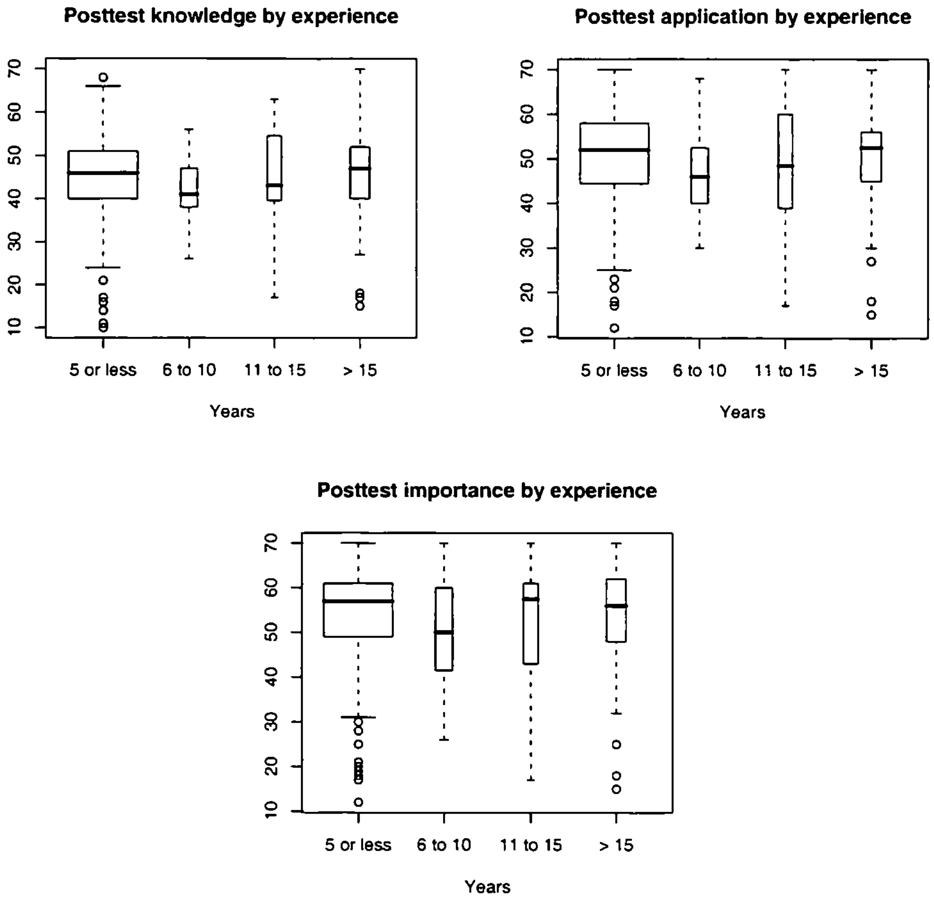


Figure 8.28: Posttest LQ subscales by *experience*, variable width.

Summary statistics of posttest LQ subscales by *experience*.

Experience (years)		*5 or less*	*6 to 10*	*11 to 15*	*> 15*
	n	638	39	28	55
Knowledge	$\bar{x}$	45.12	41.90	45.18	45.06
	s	8.66	6.28	11.09	11.20
Application	$\bar{x}$	51.02	46.85	48.61	49.96
	s	10.46	9.68	12.48	11.68
Importance	$\bar{x}$	54.93	50.97	53.96	54.00
	s	10.59	11.12	12.34	12.21

Test of differences of posttest LQ subscales by *experience*

	ANOVA			Kruskal		
	$df_{num,denom}$	F	p	df	χ^2	p
Knowledge	3, 66.75	3.03	0.04	3	7.60	0.06
Application	3, 66.61	2.54	0.06	3	7.91	0.05
Importance	3, 67.06	1.62	0.19	3	4.93	0.18

The results are similar to those of the pretest data. There is no evidence to reject the hypotheses of equal posttest Learning scores among the different experience groups.

8.9.4 Course Experience Questionnaire

With regards to the CEQ subscales, Figure 8.29 shows some variation in the medians, as well as in the interquartile ranges. In the score of the GT subscale, there appears to be a positive relationship between the score and experience. This can be an indication that those with less experience tend to be more critical of the teaching technique and approach. The same trend can also be observed in the scores of the GS subscale. Participants with 5 years or less working experience gave lower mean scores than do the other groups. This suggests that they are more critical of the skills targeted by the training programmes.

The table of summary statistics below shows the values of the mean and the standard deviation of each of the CEQ subscales by experience levels.

Summary statistics of the CEQ subscales by *experience.*

Experience	(years)	*5 or less*	*6 to 10*	*11 to 15*	> 15
	n	664	52	41	
GT	$\overline{x}$	3.39	3.64	3.70	3.89
	s	0.63	0.62	0.73	0.55
CG	$\overline{x}$	3.41	3.55	3.50	3.63
	s	0.55	0.54	0.67	0.56
GS	$\overline{x}$	3.48	3.89	3.85	3.97
	s	0.63	0.54	0.46	0.53
AA	$\overline{x}$	3.21	3.30	3.05	3.20
	s	0.60	0.61	0.77	0.68
AW	$\overline{x}$	3.11	3.13	2.90	3.43
	s	0.79	0.65	0.67	0.63

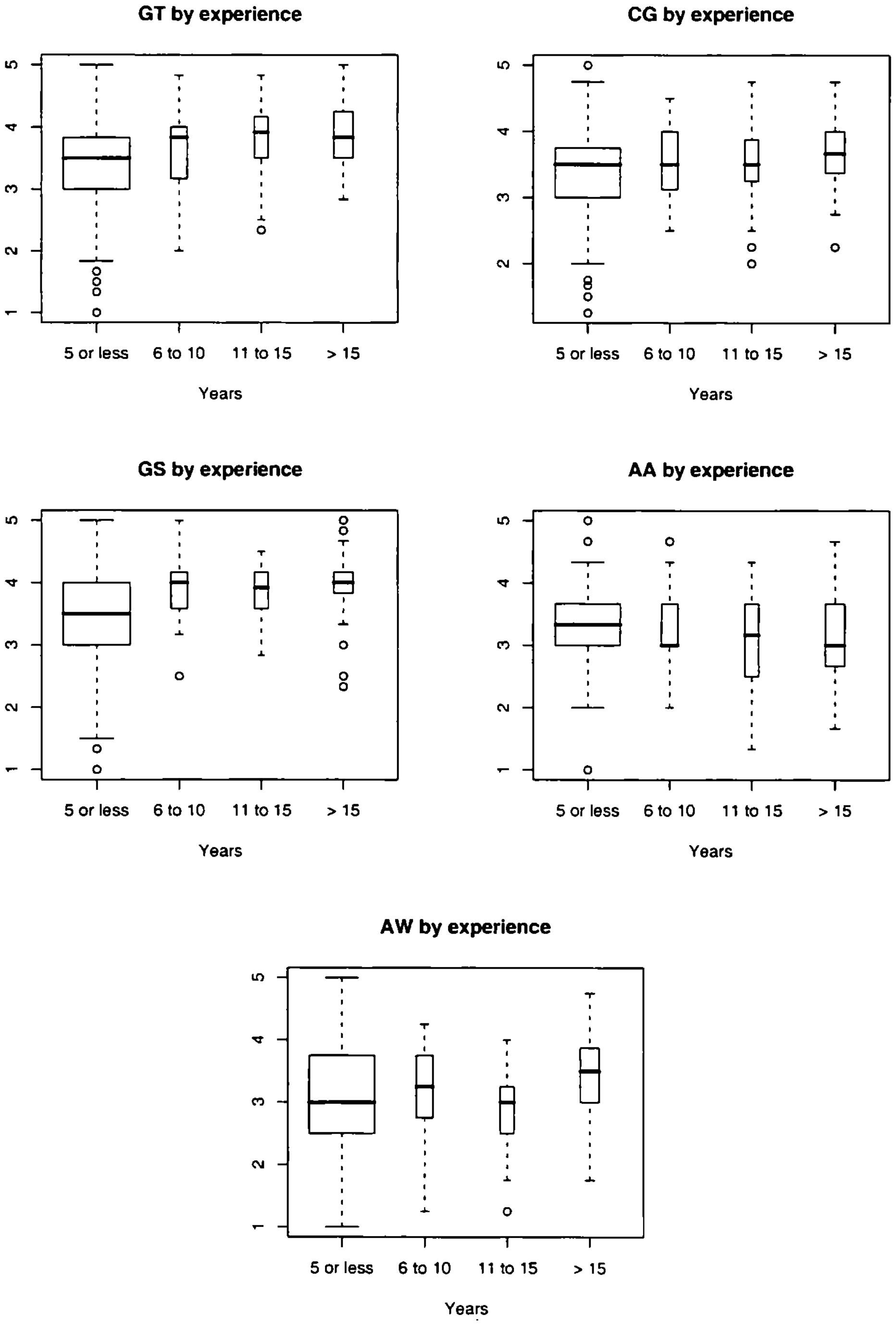


Figure 8.29: CEQ subscales by *experience*, variable width.

The Oneway ANOVA tests and Kruskal Wallis sum rank test on these data gave the following results:

Tests of differences of the CEQ subscales by *experience*.

	ANOVA			**Kruskal-Wallis**		
	$df_{num,denom}$	F	*p-value*	*df*	$KW\chi^2$	*p-value*
GT	3,67.66	15.92	6.19e-08	3	38.35	2.38e-08
CG	3,66.80	3.19	0.03	3	10.61	0.014
GS	3,70.78	22.37	2.70e-10	3	55.98	4.24e-12
AA	3,66.27	0.69	0.56	3	1.89	0.60
AW	3,70.22	5.23	0.003	3	11.96	0.008

The results show strong evidence of significant differences in the scores of the GT and GS subscales, and a weaker indication of evidence for signficant differences in the score of the AW. These results indicate that the participants with different ranges of experience perceived their training experience differently, in terms of the good teaching practice, and the important skills targeted by the training programmes.

For the GT subscale, the significant difference is observed only between the experience groups of *5 or less* and > *15 years*. For the GS, the experience group of 5 years or less has a mean score that differs significantly with the means of all the other groups. It can also be observed from the boxplots (Figure 8.29) that this group also has the largest variation among all the experience groups, indicating large amount of variability in the scores within the group itself.

8.10 Overall Conclusions

As a summary, the results of all tests of differences done in this chapter are presented in Table 8.1. From the table, several general observations can be made. First, the GHQ scale has the hypothesis of no difference rejected with strong evidence in all but two cases; no evidence of significant differences between the sexes, and a weak evidence among the ethnic groups. Secondly, the LQ and its subscales have only two instances where a significant difference are observed (though with only a weak evidence); in all other instances there are no evidence likewise. Thirdly, there is no evidence of significant differences in the subscales of the CEQ between the sexes and among all ethnic groups. For other categorical variables, the CEQ subscales produce a mixture of results, but in most cases there are more instances with evidence for significant differences than not.

Table 8.1: Strength of evidence to reject hypothesis of no difference in tests of differences.

Scales	Sub	Demographic factors						
		sex	*ethnic*	*age*	*centre*	*servSect*	*servGrp*	*exp*
GHQ		no	weak	strong	strong	strong	strong	strong
Pretest LQ	Know	no	no	no	weak	no	no	no
	App	no	no	no	no	no	no	no
	Imp	no	no	no	no	no	no	no
Posttest LQ	Know	no	no	no	no	no	no	no
	App	no	no	no	weak	no	no	no
	Imp	no	no	no	no	no	no	no
CEQ	GT	no	no	strong	strong	strong	strong	strong
	CG	no	no	weak	strong	some	strong	no
	GS	no	no	strong	strong	strong	strong	strong
	AA	no	no	no	weak	no	some	no
	AW	no	no	no	strong	some	strong	weak

Average scores of the GHQ were found to be significantly different among the different levels of the following demographic factors: *age group*, centre, *service sector*, *service group* and *experience*. These results indicate that the score of the general health is associated with the demographic factors. Since the measurement was done on the first day of training, it could not have been affected by the factors of the training itself, but most probably by other factors that had existed before training. This could include the effects of work environment, the effects of life outside work, and a lot of other possible factors. We have not however, explored differences between combinations of factors, but these too might exist. Examples are *age + sex*, *course + service sector*, etc.

In the pretest and the posttest LQ subscales, statistical tests indicated no differences among the levels of almost all demographic factors. The only exceptions are the pretest knowledge and the posttest application, both of which related to the factor of *centre*, but with only weak evidence. The general results is that none of the demographic factors are associated with the pretest and the posttest Learning.

The CEQ scale measures five different aspects of the training programmes. None of its five subscales are associated with the factor of *sex* or *ethnic groups*. We may conclude that there is no difference in the course experience between the sexes, or among the different ethnic groups. But association is more likely between the factors of the course experience subscales and the other demographic factors.

Participants from different *age groups* seem to have different views regarding the factor of GT, which regards the good teaching practice, and the factor of GS, which regards to the generic skills targetted. Results also suggest that participants attending training programmes at the different centres have different course experience, in terms of the five subcales measured.

Participants from the different *service sectors* and the different *service groups* indicate different experience in all subscales except the AA. Among the clearest difference is between the Professional staff and the Supporting staff, which suggests that the Professional staff view their training experience differently from the Supporting staff.

Participants with different levels of experience indicated different views with regards to the GT and the GS subscales. It appears that *experience* is only associated with these two factors, namely (i) their views about the good teaching practice and (ii) their views about the skills targetted by the training programme. Overall, participants from the *centre* of Finance appear to score the highest in four subscales among the centres, while results for the *centre* of IDFR indicate not much variation among the five subscales.

8.10.1 Conclusions of the Questionnaires

The GHQ

Scores of the GHQ seems to be associated with all demographic factors except *sex*, and probably *ethnic* too. Male and female respondents do not show differences in the average score of the general health. Neither do the particicants from different ethnic groups. Results suggest that average scores of the general health differ between or among the different levels of *ethnicity*, *age*, *centre*, *service sector*, *service* *group*, and *experience*.

The LQ

Scores of the subscales of Learning do not differ among all the different levels or groups of the demographic factors, except in two cases with weak evidence. The score of the pretest knowledge differs, albeit only with weak evidence, among the different *centres*. The score of the posttest application also seem to differ with weak evidence among the different *centres*. It appears that *centre* is the only demographic factor associated with Learning.

The CEQ

None of the CEQ subscales differs between the sexes or among the different ethnic groups. Both the GT and GS subscales differ significantly among the levels of *age*, *centre*, *service sector*, *service group*, and *experience*. The CG subscale is associated with *centre*, *service sector*, and *service group*. The AA subscale is associated only with *service group*. The AW subscale appear to be associated with *centre*, *service sector*, *service group* and *experience*.

Chapter 9

Differences between the Pretest and the Posttest Learning

To examine whether a training programme has any impact on Learning, measurements are made at two time points: before the programme starts (**pretest**, or **Time 1**), and after it ends (**posttest**, or **Time 2**). A successful impact by the training programme will be indicated by significant observed differences in the scores of the LQ subscales which collectively measure Learning, namely the knowledge, the application, and the importance subscales.

To be able to attribute the observed differences to the training intervention, two sets of data are used. The first is the treatment data, which was collected from participants who actually attended training. The second dataset is the control data, which was collected from a similar group of participants, but who were not attending any training during the study period. If the treatment data shows significant changes between the pretest and the posttest sets, and no changes are shown by the control data, then the idea that training impacts Learning is supported. In this chapter, the paired samples t-test is used to test the hypothesis that the mean differences between the pretest and posttest scores are zero.

Figure 9.1 shows boxplots of the differences between the posttest scores and the pretest scores for treatment and control data. The treatment data in all three subscales have wider interquartile ranges, with more outliers at both ends of the distribution, compared to the control data. All medians appear to be around zero,

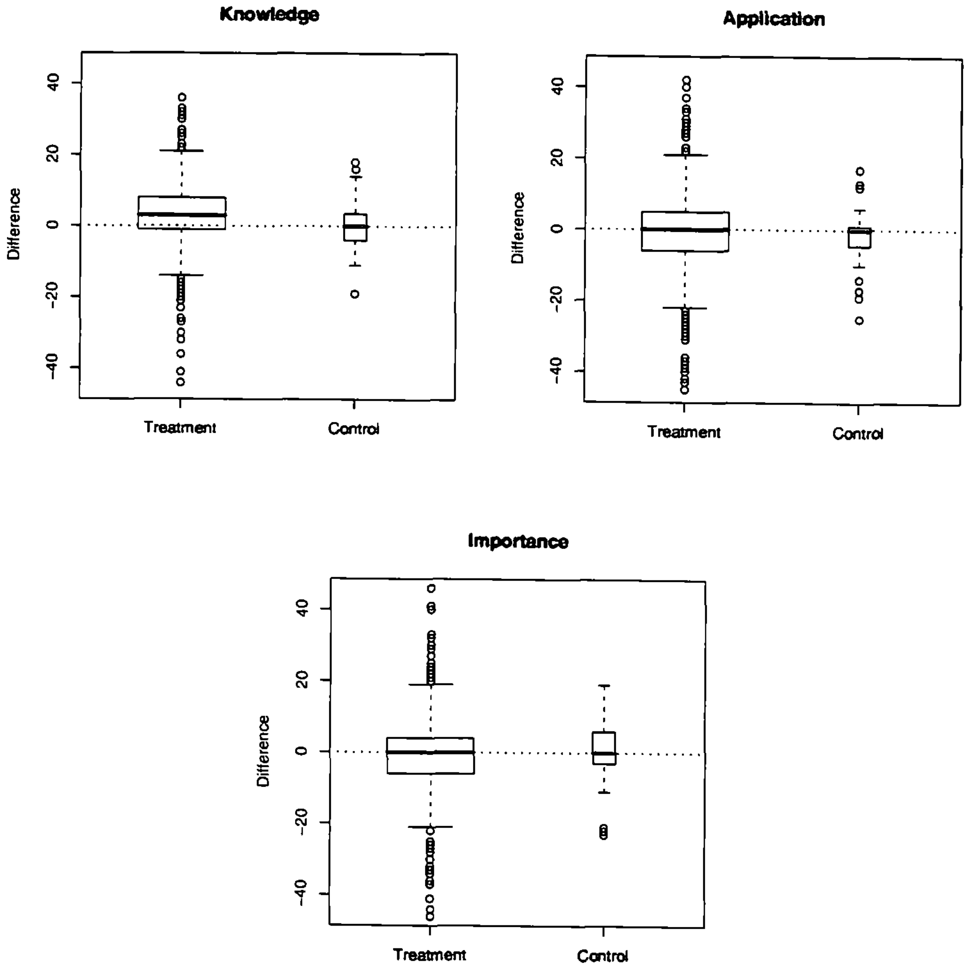


Figure 9.1: Posttest-pretest differences in Knowledge, Application, and Importance of the treatment and control data.

except for the median of the treatment knowledge. It seems to suggest a positive change in the score of knowledge among study participants that attended training. In the subscales of application and importance, no obvious difference is observed between the medians.

9.1 Statistics of the treatment and the control data.

Table 9.1 shows the statistics of the pretest and the posttest scores of the three LQ subscales of the treatment data. In the rows marked as 'Tl' and 'T2' are the observed means and standard deviations of the scores of the three LQ subscales. In the row marked 'T2-Tl diff.' are the statistics of the differences between the posttest and the prestest scores. We observe an increase in the scores of knowledge, a slight decrease in the observed scores of application, and a small decrease in the scores of importance.

Table 9.1: Summary statistics of the treatment LQ subscales

		Knowledge	*Application*	*Importance*
	n	755	755	756
Tl	$\bar{x}$	41.93	51.24	56.23
	s	8.74	10.34	10.07
T2	$\bar{x}$	44.95	50.64	54.63
	s	8.86	10.62	10.82
T2-Tl diff.	Mean of diff.	3.04	-0.60	-1.54
	Sd of diff.	9.31	11.47	11.16

The paired sample t-test is carried out on the data, to test for the hypothesis of zero mean difference between the posttest and the pretest data. Statistics from the hypothesis test are presented in Table 9.2.

The alternative non-parametric test is Wilcoxon signed rank test, which is actually a one sample test on the differences of the ranks of the scores. For these hypotheses, it produces the statistics in Table 9.3.

In both tests, the hypothesis of zero mean difference is rejected with a strong

Table 9.2: Paired-sample t-test on the T2-T1 differences of the treatment data.

	Knowledge	*Application*	*Importance*
t	8.97	-1.43	-3.79
df	754	754	755
p-value	2.2e-16	0.15	0.00016
95% CI	(2.37, 3.70)	(-1.42, 0.22)	(-2.33, -0.74)

Table 9.3: Wilcoxon signed-rank test on the T2-T1 differences of the treatment data.

	Knowledge	*Application*	*Importance*
V	175871.5	106015	79968.5
p-value	2.2e-16	0.4123	0.00082

evidence for knowledge, and with some evidence for importance. The hypothesis is not rejected for application. The average score of the posttest knowledge is different from the average score of the pretest knowledge, suggesting a change in the participants' perceived level of knowledge over the period of training. The observed scores indicate that the change is positive, an increase from the pretest to the posttest. For the application factor, the data shows no evidence that the means differ between before and after training. In other words, the findings suggest that the participants have maintained their view about the application of the subject areas in their workplaces over the training period. Meanwhile for the importance factor, the hypothesis of no difference is rejected with some evidence. The observed values suggest that the change is a decrease from the pretest to the posttest. If this apparently counter-intuitive result is true, then it means that the participants put less importance on learning the subject areas after the training, than they do before the training.

The same examination is carried out on the control data. Since the control respondents did not attend any training over the study period, there are no pretest and posttest Learning scores in the actual sense. However, data was collected twice, at timepoints T1 and T2, to mimic the pretest and the posttest among the study (treatment) respondents. However, for the whole of control data, a single T1 to T2 period of one week was used. The period of one week was chosen because it was about the average of the length of training programmes. Furthermore, having a single T1 to T2 period for the whole of control respondents greatly reduced the works involved in the distribution and the collection of the questionnaires. Table 9.4 shows the statistics of the control data.

Table 9.4: Summary statistics of the control LQ subscales

		Knowledge	*Application*	*Importance*
	n	52	51	50
Pretest	$\overline{x}$	30.36	35.11	43.46
	s	8.67	9.97	8.15
Posttest	$\overline{x}$	30.84	33.86	43.40
	s	8.17	9.87	10.29

Table 9.5: Paired-sample t-test on the T2-T1 difference of the control data.

	Knowledge	*Application*	*Importance*
t	0.51	-1.30	0.098
df	51	50	49
p-value	0.61	0.20	0.92
Mean diff.	0.48	-1.45	0.12
95% CI	(-1.41, 2.37)	(-3.70, 0.80)	(-2.35, 2.59)

The results of the paired samples t-test in Table 9.5 show no evidence of significant differences between the scores of the two time points of any of the three LQ subscales among the control respondents. This finding implies that the scores of the three LQ subscales do not change from time 1 to time 2. This supports the idea that the scores of knowledge, application, and importance do not change among those who are not attending any training. These results are further strengthened by the Wilcoxon signed-rank test as presented in Table 9.6.

Table 9.6: Wilcoxon signed-rank test on the post-pre difference of the control data.

	Knowledge	*Application*	*Importance*
V	393.5	259.5	331.5
p-value	0.74	0.17	0.79

9.2 Conclusion

Analyses on the treatment data have shown evidence that two of the LQ subscales, namely the knowledge and the importance, have different means for time 1 (pretest) and time 2 (posttest). There is no such evidence for the application subscale. These findings seem to suggest that training programmes do have an impact on the scores of the knowledge and the importance factors, but not the application factor. In other words, there is an association between attending training and the score of knowledge and importance, but there is none between attending training and the score of application. The differences as indicated in Table 9.1 are an increase of 3.02 in the knowledge score, and a decrease of 1.6 in the importance score. The scores of the LQ subscales are in the range of 10 to 70, thus these small differences might not be very promising in a practical sense. A larger difference would certainly be more meaningful in terms of the impact of the training intervention on the Learning factors.

Analyses on the control data show no significant differences between the time 1 and the time 2 scores of any of the LQ subscales. These results support the idea that changes in the scores of knowledge and application that we observe among the training participants can be attributed to the training they attended. Looking at the observed values, it may also be suggested that training is successful in increasing the perceived level of knowledge among the participants, as well as changing their attitude towards the importance of learning the subject areas, but in the wrong direction.

9.3 Exploratory analysis of subgroups

In this section, the tests of differences between the pretest and posttest scores of the LQ subscales are repeated on the subgroups of the treatment data. The subgroups are created based on the demographic factors, namely: (i) *sex*, (ii) *ethnic group*, (iii) *age group*, (iv) *centre*, (v) *service sector*, (vi) *service group* (vii) *experience*, and (viii) *length of training*. The purpose is to explore whether the impact of training is stronger in certain groups of participants than in the others. The tests of differences used are the paired-samples t-test, and the non-parametric alternative, the Wilcoxon signed rank test for paired samples. A more sophisticated test, in the form of the Analysis of Covariance (ANCOVA) will be utilized on the same hypotheses in the following section.

9.3.1 Sex

In this section we examine whether the changes in the LQ subscales differ between the male and female respondents. A total of 424 males and 334 females make up the respondents for this test. The corresponding percentages are about 56% and 44%. The results of the tests of differences are as in Table 9.7.

Table 9.7: Results of T2-T1 tests of differences by *sex*.

Sex	Statistics	*Knowledge*	*Application*	*Importance*
Male	P-value(T-test)	3.44e-08	0.30	0.008
	95% CI	(1.7, 3.5)	(-1.6, 0.5)	(-2.6, -0.4)
	df	420	420	421
	Mean difference	2.62	-0.57	-1.48
	P-value(Wilcoxon)	9.6e-11	0.48	0.01
Female	P-value(T-test)	4.79e-12	0.30	0.007
	95% CI	(2.5, 4.5)	(-1.9, 0.6)	(-2.8, -0.5)
	df	331	331	331
	Mean difference	3.51	-0.66	-1.62
	P-value(Wilcoxon)	2.18e-14	0.66	0.02

The hypothesis of zero mean difference in the knowledge factor is rejected in both male and female groups. There is no evidence to indicate differences in the application subscale, and there is some evidence likewise in both groups for the importance subscales. The effect is about the same for both male and female subgroups.

9.3.2 Ethnic Group

In this section we examine pretest-posttest changes in the scores of the LQ subscales among the *ethnic groups*, namely the Malays, the Indians, the Chinese and Others. Table 9.8 shows the results of the tests of differences.

Table 9.8: Results of T2-T1 tests of differences by *ethnic group*.

Ethnic	**Statistics**	*Knowledge*	*Application*	*Importance*
Malay	P-value(T-test)	1.13e-15	0.41	0.0002
	95% CI	(2.4, 3.8)	(-1.3, 0.5)	(-2.5, -0.7)
	df	642	642	644
	Mean difference	3.09	-0.37	-1.63
	P-value(Wilcoxon)	2.2e-16	0.81	0.001
Chinese	P-value(T-test)	0.003	0.49	0.94
	95% CI	(1.3, 5.7)	(-4.3, 2.1)	(-3.1, 2.9)
	df	34	34	34
	Mean difference	3.51	-1.11	-0.11
	P-value(Wilcoxon)	0.03	0.80	0.68
Indian	P-value(T-test)	0.34	0.09	0.14
	95% CI	(-1.3, 3.8)	(-6.7, 0.5)	(-6.2, 0.9)
	df	47	48	47
	Mean difference	1.23	-3.08	-2.65
	P-value(Wilcoxon)	0.26	0.08	0.16
Others	P-value(T-test)	0.005	0.87	0.70
	95% CI	(1.6, 7.9)	(-5.7, 4.8)	(-3.7, 5.4)
	df	26	25	25
	Mean difference	4.74	-0.42	0.85
	P-value(Wilcoxon)	0.006	0.65	1.0

The results of the tests of differences on the knowledge and importance subscales are not similar among all ethnic groups. Changes in the score of knowledge seems to happen in the ethnic group of Malay only. The same may be said about the score of importance, but the evidence is weaker. There is no evidence to indicate changes in the score of application in all ethnic groups.

9.3.3 Age

Changes between the pretest and the posttest of the LQ subscales may not occur similarly in all age groups, namely (i) <26 years, (ii) 26-30 years, (iii) 31-35 years, (iv) 36-40 years, (v) 41-45 years, and (vi) >45 years. The results of the T1-T2 tests of differences are in Table 9.9, where each group is represented by its middle value, called the midpoint.

Plots in Figure 9.2 show the mean differences and the confidence intervals on the hypothesis that the mean differences are zero. The plots suggest that the average

Table 9.9: Results of T2-T1 tests of differences by *age group*.

Age midpoint	Statistics	*Knowledge*	*Application*	*Importance*
23 yrs.	P-value(T-test)	5.4e-09	0.37	0.002
	95% CI	(2.7, 5.3)	(-2.02, 0.75)	(-3.5, -0.8)
	df	184	181	184
	Mean difference	4.0	-0.64	-2.14
	P-value(Wilcoxon)	1.5e-09	0.62	0.007
28 yrs.	P-value(T-test)	8.3e-10	0.11	0.02
	95% CI	(1.9, 3.6)	(-2.05, 0.22)	(-2.3, -0.2)
	df	354	356	354
	Mean difference	2.76	-0.91	-1.27
	P-value(Wilcoxon)	6.7e-12	0.38	0.06
33 yrs.	P-value(T-test)	0.03	0.46	0.09
	95% CI	(0.3, 4.0)	(-3.0, 1.4)	(-3.8, 0.3)
	df	127	127	127
	Mean difference	2.12	-0.80	-1.75
	P-value(Wilcoxon)	0.01	0.42	0.09
38 yrs.	P-value(T-test)	0.01	0.85	0.36
	95% CI	(1.2, 8.3)	(-4.9, 5.9)	(-8.5, 3.1)
	df	29	29	29
	Mean difference	4.77	0.5	-2.67
	P-value(Wilcoxon)	0.006	0.63	0.42
43 yrs.	P-value(T-test)	0.56	0.30	0.88
	95% CI	(-4.4, 7.9)	(-3.1, 9.8)	(-6.5, 7.6)
	df	28	28	28
	Mean difference	1.79	3.34	0.52
	P-value(Wilcoxon)	0.32	0.23	0.67
52 yrs.	P-value(T-test)	0.02	0.99	0.75
	95% CI	(0.9, 8.2)	(-6.3, 6.2)	(-6.8, 5.0)
	df	24	25	25
	Mean difference	4.52	-0.04	-0.92
	P-value(Wilcoxon)	0.01	0.99	0.84

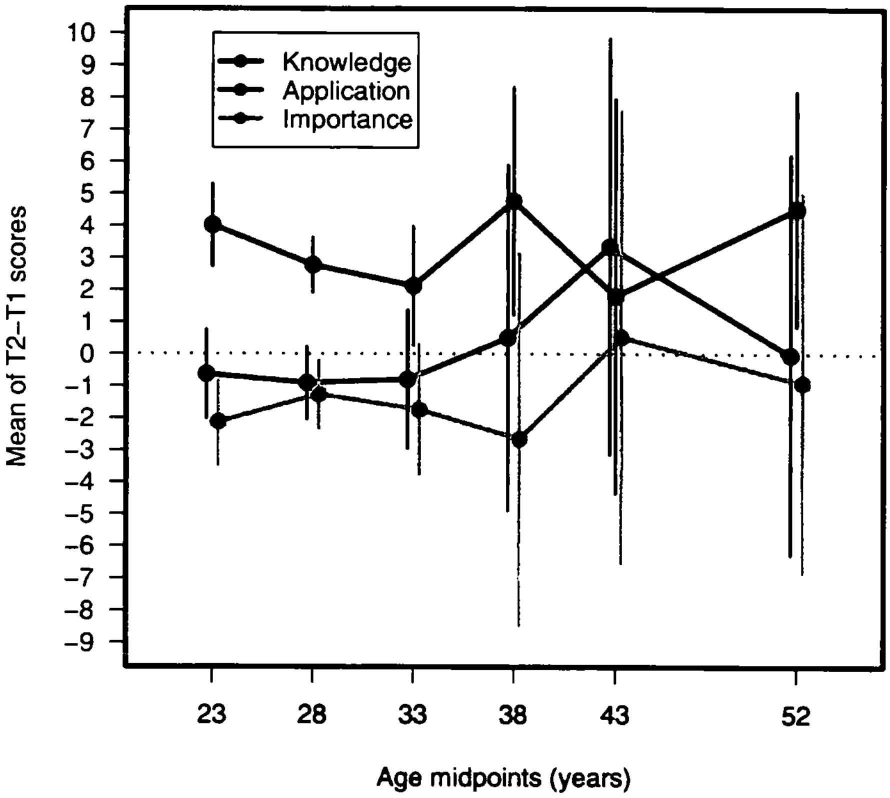


Figure 9.2: Mean differences and their respective 95% confidence intervals of the T2-T1 differences in the three LQ subscales by age groups.

score of knowledge increases from before the training to after the training in the first two groups, namely the group with midpoints of 23 and 28. For the subscale of application, there is no indication that changes in the scores occur in any age groups. For the subscale of importance, there is a slight evidence that changes in the score occurs in the first two groups, namely the group with midpoint 23 and the group with midpoint 28.

9.3.4 Centre

The impact of training on the participants should ideally occur in all programmes from all centres of INTAN. In this section we examine whether that is the case. The results of the tests of differences by *centre* are in Table 9.10.

Table 9.10: Results of T2-T1 tests of differences by *centre*.

Centre	Statistics	*Knowledge*	*Application*	*Importance*
Mgt	(P-value(T-test)	0.12	0.49	0.62
	95% CI	(-1.2, 10.0)	(-4.3, 8.7)	(-4.5, 7.5)
	df	25	25	25
	Mean difference	4.38	2.19	1.5
	P-value(Wilcoxon)	0.03	0.34	0.28
Economic	P-value(T-test)	4.6e-08	0.004	0.004
	95% CI	(3.2, 6.6)	(-4.9, -0.9)	(-4.3, -0.8)
	df	153	153	153
	Mean difference	4.95	-2.94	-2.55
	P-value(Wilcoxon)	2.6e-09	0.007	0.02
KTD	P-value(T-test)	0.15	0.82	0.93
	95% CI	(-3.2, 16.7)	(-15.7, 19.2)	(-8.1, 7.5)
	df	7	6	6
	Mean difference	6.75	1.7	-0.29
	P-value(Wilcoxon)	0.14	0.35	0.99
Quantitative	P-value(T-test)	0.06	0.02	0.0003
	95% CI	(-0.1, 3.2)	(-4.6, -0.4)	(-5.9, -1.8)
	df	168	169	170
	Mean difference	1.58	-2.48	-3.87
	P-value(Wilcoxon)	0.02	0.02	0.0004
Imatec	P-value(T-test)	8.7e-06	0.35	0.91
	95% CI	(1.6, 3.9)	(-0.8, 2.1)	(-1.5, 1.6)
	df	154	154	153
	Mean difference	2.72	0.69	0.09
	P-value(Wilcoxon)	9.5e-06	0.53	0.92
Finance	P-value(T-test)	0.13	0.58	0.15
	95% CI	(-1.0, 7.4)	(-2.4, 4.2)	(-8.1, 1.3)
	df	27	27	27
	Mean difference	3.18	0.89	-3.39
	P-value(Wilcoxon)	0.02	0.53	0.21
IDFR	P-value(T-test)	9.6e-09	0.08	0.63
	95% CI	(1.8, 3.6)	(-0.1, 2.2)	(-1.5, 0.9)
	df	214	214	215
	Mean difference	2.73	1.03	-0.29
	P-value(Wilcoxon)	1.3e-09	0.02	0.65

Results of the tests indicate that the scores have changed substantially in programmes from the centre of Economic, the centre of Imatec and the centre of IDFR, for the knowledge subscale. All of them indicate an increase from before the training to after the training. There is no strong evidence to support changes in the scales of application and importance.

9.3.5 Service Sector

In this section we examine the association between the pretest-posttest difference of the LQ subscales and *service sector*. The majority (664) of the study respondents work with the Federal Government. Another 52 work with the State Governments, and the rest (41) work with the Local Governments from all over the country. Table 9.11 shows the results of the T2-T1 tests of differences by *service sector*.

Table 9.11: Results of the T2-T1 tests of differences by *service sector*.

Service Sector	Statistics	*Knowledge*	*Application*	*Importance*
Federal	P-value(T-test)	3.13e-14	0.27	0.0004
	95% CI	(2.1, 3.5)	(-1.4, 0.4)	(-2.4, 0.7)
	df	658	659	660
	Mean difference	2.82	-0.49	-1.56
	P-value(Wilcoxon)	2.2e-16	0.65	0.001
State	P-value(T-test)	0.0003	0.27	0.26
	95% CI	(2.6, 8.1)	(-4.8, 1.4)	(-4.8, 1.3)
	df	51	50	50
	Mean difference	5.37	-1.72	-1.73
	P-value(Wilcoxon)	2.58e-05	0.28	0.30
Local	P-value(T-test)	0.0007	0.96	0.53
	95% CI	(1.9, 6.6)	(-3.0, 3.1)	(-4.4, 2.3)
	df	40	40	40
	Mean difference	4.29	0.07	-1.05
	P-value(Wilcoxon)	0.001	0.98	0.81

Generally there is evidence of significant differences between the pretest and the posttest knowledge in all three sectors. Thus there is no obvious association between the pretest-posttest differences in knowledge and the *service sector*. For the application subscale, there is no evidence to suggest significant differences in all three service sectors. For the importance subscale, Federal is the only service group for which there is some evidence of a significant difference. In this same subscale, the subgroup of State shows a larger mean difference than that of the Federal, but probably the sample size is inadequate to detect genuine changes.

9.3.6 Service Group

In this section, we examine whether the differences between the pretest and the posttest scores of the LQ subscales are associated with *service group*. There are two service groups, namely the Professional and Management group and the Supporting Staff group. A total of 708 respondents are from the Professional and Management group while only 47 are from the Supporting Staff.

Table 9.12: Results of T2-T1 tests of differences by *service group*.

Group	Statistics	*Knowledge*	*Application*	*Importance*
Prof	P-value(T-test)	2.31e-16	0.05	7.16e-05
	95% CI	(2.3, 3.6)	(-1.7, -0.006)	(-2.5, -0.8)
	df	702	702	703
	Mean difference	2.96	-0.86	-1.66
	P-value(Wilcoxon)	2.2e-16	0.16	0.0004
Supp	P-value(T-test)	0.01	0.02	0.76
	95% CI	(0.9, 6.3)	(0.5, 6.1)	(-3.2, 4.4)
	df	46	46	46
	Mean difference	3.62	3.30	0.57
	P-value(Wilcoxon)	0.0009	0.03	0.57

Table 9.12 shows the results of the tests. It appears that the scores are different for knowledge and importance for the Professional group. There is no evidence likewise for the Supporting group.

9.3.7 Experience

The *experience* factor relates to the length of time (in the range of years) the study participants have been working in the Malaysian public service. In this section we explore whether T2-T1 changes in the scores of the LQ subscales occur in subgroups of different experience.

The results of the tests of differences in Table 9.13, as well as the graphical presentation in Figure 9.3, suggest that changes do occur in some experience groups. For the knowledge subscale, there is evidence that scores increase in the group with $\leq$5 years experience (midpoint=2.5). There is no clear indication for the subscale of application. For the importance subscale, the score is likely to decrease in the

Table 9.13: Results of T2-T1 tests of differences by *experience*.

Experience midpoint	**Statistics**	*Knowledge*	*Application*	*Importance*
2.5 yrs.	P-value(T-test)	2.2e-16	0.16	0.0006
	95% CI	(2.5, 3.9)	(-1.4, 0.2)	(-2.2, -0.6)
	df	635	634	635
	Mean difference	3.20	-0.6	-1.4
	P-value(Wilcoxon)	2.2e-16	0.47	0.002
8 yrs.	P-value(T-test)	0.83	0.04	0.008
	95% CI	(-3.2, 2.6)	(-8.6, -0.1)	(-9.8, -1.6)
	df	38	38	38
	Mean difference	-0.31	-4.36	-5.7
	P-value(Wilcoxon)	0.71	0.04	0.01
13 yrs.	P-value(T-test)	0.36	0.32	0.18
	95% CI	(-2.6, 6.8)	(-8.6, 2.9)	(-8.6, 1.7)
	df	27	27	27
	Mean difference	2.11	-2.8	-3.46
	P-value(Wilcoxon)	0.04	0.46	0.30
20 yrs.	P-value(T-test)	0.02	0.11	0.57
	95% CI	(0.7, 7.4)	(-0.8, 7.6)	(-3.1, 5.6)
	df	51	52	52
	Mean difference	4.04	3.4	1.23
	P-value(Wilcoxon)	0.004	0.11	0.38

first experience groups, ie. those participants with ≤5 years experience, and with a much less evidence, among those with 6-10 years experience.

9.3.8 Length of Training

The majority of training programmes in INTAN are run between 2 to 10 days. The respondents of this study attended programmes which ran for five different number of days, namely 3, 4, 5, 9 and 10 days. The distribution of respondents by the number of days of the training programmes are as follows:

Length (days)	*3*	*4*	*5*	*9*	*10*
Number of respondents	25	48	515	70	102

Table 9.14: Results of T2-T1 tests of differences by *length of training*.

Length	**Statistics**	*Knowledge*	*Application*	*Importance*
3 days	P-value(T-test)	0.02	0.25	0.46
	95% CI	(1.2, 10.7)	(-2.9, 10.7)	(-3.7, 7.8)
	df	24	23	23
	Mean difference	5.96	3.92	2.08
	Wilcoxon	0.01	0.15	0.42
4 days	P-value(T-test)	0.05	0.50	0.22
	95% CI	(0.04, 5.2)	(-1.5, 3.0)	(-4.8, 1.2)
	df	47	47	47
	Mean difference	2.63	0.8	-1.83
	P-value(Wilcoxon)	0.007	0.54	0.45
5 days	P-value(T-test)	2.2e-16	0.50	0.04
	95% CI	(2.7, 4.2)	(-1.2, 0.6)	(-1.8, -0.03)
	df	512	512	512
	Mean difference	3.42	-0.31	-0.90
	P-value(Wilcoxon)	2.2e-16	0.93	0.09
9 days	P-value(T-test)	0.45	0.88	0.62
	95% CI	(-1.7, 3.7)	(-3.7, 4.3)	(-4.7, 2.8)
	df	69	69	69
	Mean difference	1.03	0.31	-0.93
	P-value(Wilcoxon)	0.29	0.93	0.59
10 days	P-value(T-test)	0.07	0.0002	1.9e-06
	95% CI	(-0.2. 4.1)	(-6.7, -2.2)	(-8.2, -3.6)
	df	98	99	100
	Mean difference	1.97	-4.44	-5.9
	P-value(Wilcoxon)	0.04	0.001	8.9e-06

The results in Table 9.14 suggest a significant change in the score of knowledge for the 5 day programmes. Change in the score of importance appears to occur only for the 10 day programmes. In the same group, there is also a slight evidence to indicate change in the score of application.

9.3.9 Summary of the Pretest-Posttest Differences by Demographic Factors

Table 9.15 shows the levels of the demographic factors where the tests of differences display evidences for significant differences between the pretest and the posttest scores of each of the LQ subscales.

Table 9.15: Levels of the demographic factors where there are evidences of significant T2-T1 differences for each of the LQ subscales.

Factors	**Knowledge**	**Application**	**Importance**
Sex	Male Female		
Ethnic	Malays		Malays
Age	<26 26-30		<26
Centre	Economic IMATEC IDFR		Quantitative
Service sector	Federal State Local		Federal
Service group	Professional		Professional
Experience	≤5 yrs		≤5 yrs
Length	5 days	10 days	10 days

Tests of differences using paired-samples t-test in this section is exploratory, and the Analysis of Covariances (ANCOVA) in the following section is a better method to test the same hypotheses. Nevertheless, we will have a brief look at what the results of this section suggest.

Overall, changes from the pretest to the posttest score seem to occur more often for the subscale of knowledge than for the other two subscales. In fact, there is almost no indication for a significant change for the application subscale. Among the demographic factors, *sex* and *service sector* are the only two where changes in knowledge happen in all of their levels. This implies that *sex* and *service sector* are not associated with the T2-T1 changes in knowledge. In all other factors, it is plausible that changes in the average scores of the knowledge and importance happen in certain subgroups only.

9.4 Using ANCOVA for Estimating Treatment and Subgroup Effects

In Chapter 8, we carried out t-test and ANOVA to test whether the average scores of scales differ between or among the different levels of the demographic factors. The general finding for the Learning data in that chapter is that the average scores of the pretest and the posttest LQ subscales do not differ between or among the different levels of the seven demographic factors. In this section, we will be using the ANCOVA to test the same hypothesis, but this time with a different methodology and focusing just on the Learning data.

The ANCOVA is a general linear model, where a covariate is used to control for the initial differences among the participants in the study. Stephen Senn [81] strongly makes the case that ANCOVA is the right method to use in cases like this study, where there are pretest and posttest scores and demographic factors. The pretest score is taken to be the covariate, the posttest score is the dependent variable, and the factor levels are the treatments to be tested.

With ANCOVA, the researcher is able to answer what would happen to the posttest scores if all participants score equally on the pretest score. The pretest score is a good covariate even if it does not differ significantly among the groups, so long as its correlation with the dependent variable is large [85].

Nunnally and Bernstein [67] explain that the main and the most appropriate use of ANCOVA is when the covariate and the criterion are highly correlated, but the subjects are assigned at random. This way ensures that the covariate and treatment effects are uncorrelated. In the hierarchical (incremental) approach to eliminate the estimated effects of variables of lesser interest, the covariate is entered first or before treatment effects of more focal interest. ANCOVA also helps to reduce within-group variability, which is due primarily to individual differences among study participants.

9.4.1 How ANCOVA Works

ANCOVA adjusts the group means of the dependents to what they would be if all groups started out equally on the covariate. The groups are adjusted to the overall means. In a pretest-posttest study with the posttest score as the dependent variable and the pretest score as the covariate, the posttest means of all the groups are adjusted to be the values they would be if all groups had started on equal pretest scores.

The relationship between the covariate and the dependent variable is assumed to be linear, for all groups. Moreover, the slopes of the regression lines of the association between the covariate and the dependent variable for all groups are also assumed to be equal. This assumption is on the population slopes, and not on the sample slopes. The analysis can only proceed if the sample slopes do not differ too much to conclude that the population slopes are not equal.

Scatterplots of the covariate with the dependent variable for each of the groups can give an indication as to how similar the slopes of the groups are, but for small sample sizes visual inspection is not reliable [85]. This is the situation for many of the cases in this study. There are many groups where the observations are small, thus graphical examination on the scatterplots is not easy to interpret.

ANCOVA will reduce the amount of total variability in the dependent variables (all groups combined) by as much as the coefficient of determination (r^2) between the dependent variable and the covariate. Once that part of variance is removed, focus is turned to the 'residual variance'. As a consequence of this removal, the within-groups variability of the dependent variable will decrease. An F-test that depends on the ratio of between-group variability to within-group variability has therefore increased in power. The use of a covariate thus can make a difference between not finding significance, and finding the genuine effects which are not obscured by the presence of a covariate.

The final test of ANCOVA is similar to the F-test in ANOVA, where the main hypothesis to be tested is the equality of the means. The only difference is that it uses adjusted values of the **sum of squares between groups** (SS_{bg}) and the **sum of squares within groups** (SS_{wg}), as well as adjusted degrees of freedom for SS_{wg}. The df (in ANOVA it is $N - k$) is reduced by one to become $N - k - 1$, to accommodate the removal of the covariance portion of the variability [60].

Now the F-test for the adjusted means becomes:

$$F = \frac{MS_{bg}}{MS_{wg}} = \frac{SS_{bg}/df_{bg}}{SS_{wg}/df_{wg}} = \frac{SS_{bg}/k-1}{SS_{wg}/N-k-1}$$

The test is not testing for a significant difference between the original groups means, but testing on the adjusted means, which usually are different from the originals. Lowry [60] explains that the conclusions to be drawn from the F-test are not as straightforward as for a normal ANOVA, but are tied together by the following statements:

- that the correlation between dependent variable and covariate within the general population is approximately the same as we have observed within the samples;
- that we remove from dependent variable the covariance that it has with the covariate, so as to remove from the analysis the pre-existing individual differences that are measured by the covariate; and
- that we adjust the group means of the dependent variable in accordance with the observed correlation between the covariate and the dependent variable.

9.4.2 Coding of Independent Variables

Predictor variables can be coded in many different ways. Faraway ([25] and [26]) suggests a few ways of coding qualitative predictors using dummy coding. The choice of coding does not affect the r^2, σ^2 and overall F statistic, but it does affect the regression coefficients. Another way of using the categorical variables is to make them factors [59]. In the R software, this is done by using the code '`factor(x)`' before starting on ANCOVA analysis. This is the approach used for the analysis in this study.

When the factors are compared among the levels, R uses *treatment coding* by default [25]. In this type of coding, level one is treated as the standard level to which all other levels are compared and referred. For all analyses in this section however, *sum coding* is used instead. Using this coding, the coefficients sum up to zero, making examination easier. Nevertheless, the coefficient of one of the levels is still not presented in the output.

9.4.3 Analysis of Covariance on the Learning Data.

In this analysis, each of the LQ subscales (knowledge, application and importance) is tested. Each time, the subscale's pretest score (T1) is the covariate while the posttest score (T2) is the dependent variable. The basic model tested in this analysis is as follows:

$$\textit{Dependent variable} = \beta_0 + \beta_1(\textit{Covariate}) + \varepsilon$$

where β_0 and β_1 are the regression coefficients, and ε is the error. This basic model is called **Model 1**, and it is the first model tested for each scale. In each of the models following Model 1, a demographic factor is included in the model as a predictor variable, as well as the interaction term of the factor variable and the covariate. The demographic factors are *sex*, *age group*, *ethnic group*, *service group*, *service sector*, *centre* and *experience*. With the factor and the interaction term in, the model becomes:

$$\textit{Dependent variable} = \beta_o + \beta_1(\textit{Covariate}) + \beta_2(\textit{Predictor}) + \beta_3(\textit{Covariate}*\textit{Predictor}) + \varepsilon$$

where β_2 is the regression coefficient for the predictor variable, and β_3 is the regression coefficient for the interaction term. The hypothesis tested is whether the predictor variable significantly predicts the dependent variable, after controlling for the variation in the covariate. If sufficient evidence is found to conclude so, it suggests that the average value of the dependent variable (the posttest score) differs between or among the different levels of the predictor variable, after adjusting for the covariate (the pretest score). In other words, the predictor variable is able to explain the remaining variation in the dependent variable. The other hypothesis tested involves the interaction term; if it is found to be significant, difference in slopes among the factor levels is suggested, which means different levels of pretest-posttest association between among the levels.

Besides the seven demographic factors as listed above, we also examine the association between the posttest scores of the LQ subscales with (i) *course* and (ii) the GHQ score. *Course* is an additional factor of interest, and we would like to know whether *course* is related to posttest Learning. We also would like to know whether Learning is associated with the state of the psychological health of the participants, thus the inclusion of the GHQ score in this analysis.

Naturally, results of the analysis are presented in two types of tables, the **ANOVA table** and the **table of coefficients**. In this chapter, the table of coefficients are only included for cases of significant ANOVA. In the tables, **T1** means *time 1* or the pretest, while **T2** means *time 2* or the posttest. The scatterplots of the factor's pretest-posttest scores are also presented, to give indications of the regression slopes. The plot of the first level is at the bottom-left, the second one is to its right, and so on up to as appropriate, as indicated in the following example of 6 levels:

Level 4	Level 5	Level 6
Level 1	Level 2	Level 3

Some of the scatterplots in the following analyses show scores of the pretest and posttest which are maximum, or very close to the maximum, which is 70. This applies to all three scales, namely the knowledge, the application, and the importance. They do not receive special analysis, as it is hard to provide statistical methods which can comfortably handle such 'ceiling' effects.

9.4.4 Knowledge Subscale

In this section, the posttest knowledge score is the dependent variable while the pretest knowledge score is the covariate.

Table 9.16: Test results of Knowledge basic model.

ANOVA for model 1

	df	SS	F	Pr(>F)
Know T1	1	10921	174.80	2.2e-16
Resid	747	46671		

Coefficients:

	Est	*SE*	t	Pr(> $|t|$)
Intercept	26.50	1.43	18.52	<2e-16
Know T1	0.44	0.03	13.22	<2e-16

Model 1

The result of this model is in Table 9.16. It suggests that the pretest knowledge score is a significant predictor variable for the posttest knowledge score. The intercept is 26.5, a value suggested to be the value of the posttest knowledge score without the pretest input. An increase of one unit in the pretest score would result in an increase of 0.44 unit in posttest knowledge score. The correlation coefficient between them is 0.435, giving an R^2 value of just about 18.96%.

Model 2 : Sex

The first demographic factor included as a predictor variable is *sex*. Figure 9.4 indicates that the distribution of pretest knowledge score against posttest score appears to be more spread out for male as compared to female. However, results of the F test on the interaction in Table 9.17 suggests there is no difference in the slopes of male and female respondents. They also suggest that *sex* does not explain the variation in posttest knowledge score, after adjusting for pretest knowledge score. It means that the average posttest knowledge score does not differ between male and female respondents.

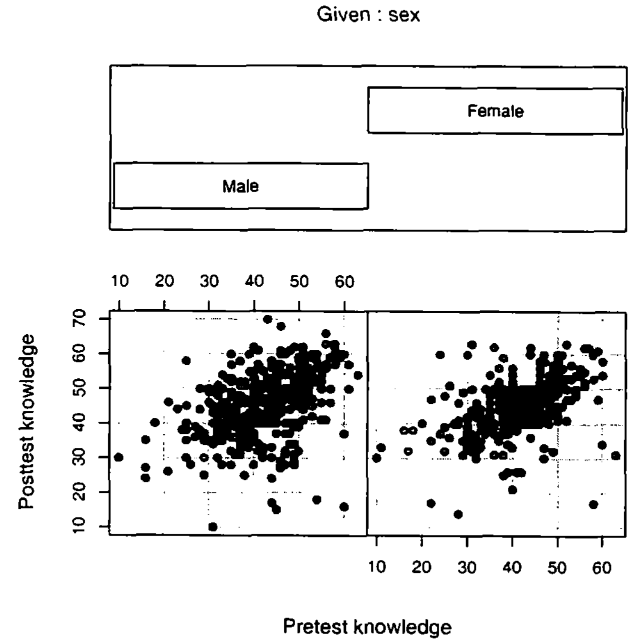


Figure 9.4: Knowledge by *sex*.

Table 9.17: Knowledge predicted by *sex*.

ANOVA for model 2

	df	SS	F	Pr(>F)
Know T1	1	11076	177.29	2e-16
Sex	1	9	0.14	0.71
Know T1*Sex	1	10	0.16	0.69
Resid	743	46472		

Model 3 : Age Group

Figure 9.5 suggests that there might be differences in the regression slopes among the age groups. The results of the test in Table 9.18 suggest that after controlling for the pretest knowledge score, *age group* does not appear to be a strongly significant predictor variable of the posttest knowledge score. The posttest score does not differ among the different age groups, after adjusting for the pretest score.

Model 4 : Ethnic Group

In model 4, the demographic factor tested as a predictor variable is *ethnic group*. Figure 9.6 shows no evidence to suggest significant differences among the ethnic

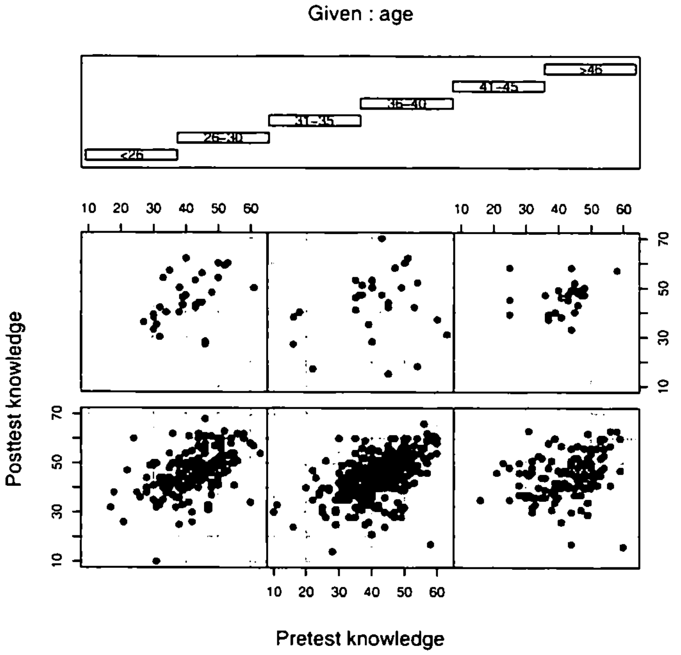


Figure 9.5: Knowledge by *age group*.

Table 9.18: Knowledge predicted by *age group*.
ANOVA for model 3

	df	SS	F	Pr(>F)
Know T1	1	10825	174.95	<2e-16
AgeGroup	5	416	1.34	0.24
Know T1*AgeGroup	5	739	2.39	0.04
Resid	739	46155		

groups. The test results in Table 9.19 indicate that *ethnic group* is not a significant predictor variable of the posttest knowledge score, after adjusting for the pretest score.

Model 5 : Service Sector

For model 5, *service sector* factor is the predictor variable. The plots in Figure 9.7 do not indicate evidence of differences in the slopes or the intercepts. However, a ceiling effect is obvious in the Federal level. The results in Table 9.20 indicate, with a weak evidence, that *service sector* is related to the posttest knowledge score, when the pretest knowledge score is adjusted for. It suggests that where participants work,

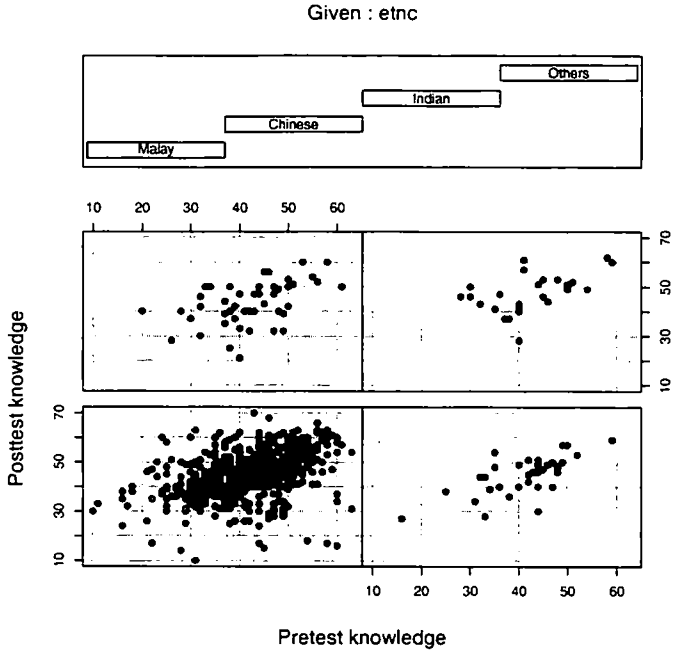


Figure 9.6: Knowledge by *ethnic group.*

Table 9.19: Knowledge predicted by *ethnic group.*
ANOVA for model 4

	df	SS	F	Pr(>F)
Know T1	1	10887	174.25	<2e-16
Ethnic	3	297	1.58	0.19
Know T1*Ethnic	3	131	0.70	0.55
Resid	739	46172		

whether in the Federal Agencies, State Governments, or Local Authorities, matters very slightly in their evaluation of the posttest knowledge.

Model 6 : Service Group

The next demographic factor is *service group*, referring to the two general groups of participants, namely the Management and Professional (Prof), and the Supporting group (Support). Figure 9.8 suggests no evidence to indicate differences. The test results in Table 9.21 shows no significant difference in the means of the posttest knowledge between the Professional officers and the Supporting staff, after pretest knowledge score is controlled for.

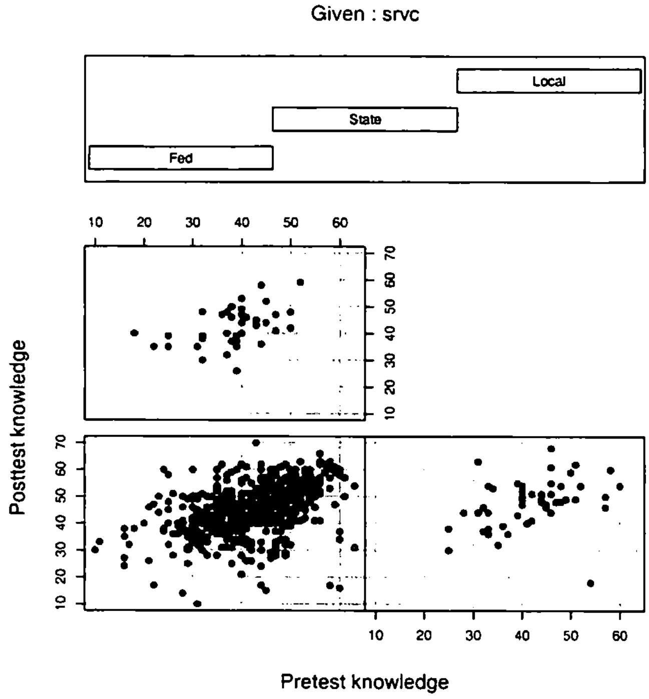


Figure 9.7: Knowledge by *service sector*.

Table 9.20: Knowledge predicted by *service sector*. ANOVA for model 5

	df	SS	F	Pr(>F)
Know T1	1	10893	175.59	<2e-16
Service	2	328	2.64	0.07
Know T1*Service	2	18	0.15	0.86
Resid	740	45905		

Table 9.21: Knowledge predicted by *service group*. ANOVA for model 6

	df	SS	F	Pr(>F)
Know T1	1	10532	167.84	<2e-16
Group	1	11	0.17	0.68
Know T1*Group	1	28	0.45	0.50
Resid	740	46435		

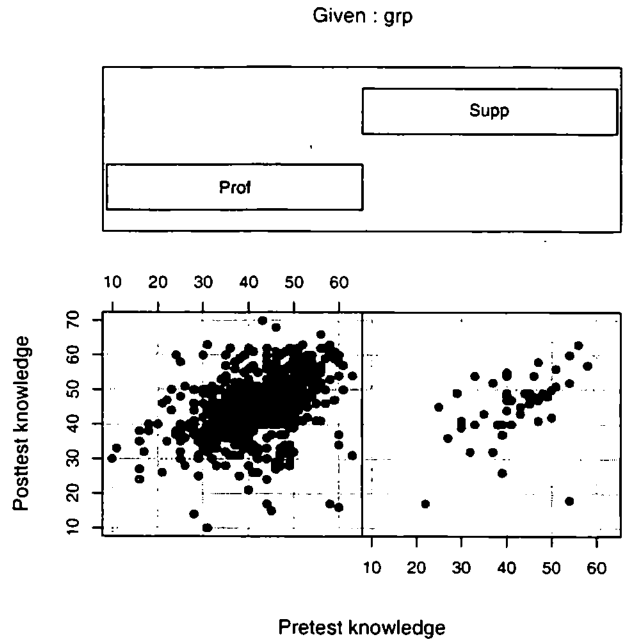


Figure 9.8: Knowledge by *service group.*

Table 9.22: Knowledge predicted by *experience.*

ANOVA for model 7

	df	SS	F	Pr(>F)
Know T1	1	10921	175.70	<2e-16
Experience	3	442	2.37	0.07
Know T1*Experience	3	169	0.90	0.44
Resid	741	46061		

Model 7 : Experience

In model 7, *experience* is the demographic factor tested as a predictor variable. Figure 9.9 indicates that the distribution of the experience group of 5 or less is more clustered compared to the other groups. The slope of the 11 to 15 and 15 or more groups are also not very clear.

The results in Table 9.22 shows that there is only a slight evidence to suggest that *experience* is significant in predicting the posttest knowledge score, after adjusting for the pretest score. Looking at the plots, samples sizes are small for the last three groups, so we decided to combine them and re-analyze. Still we find no evidence

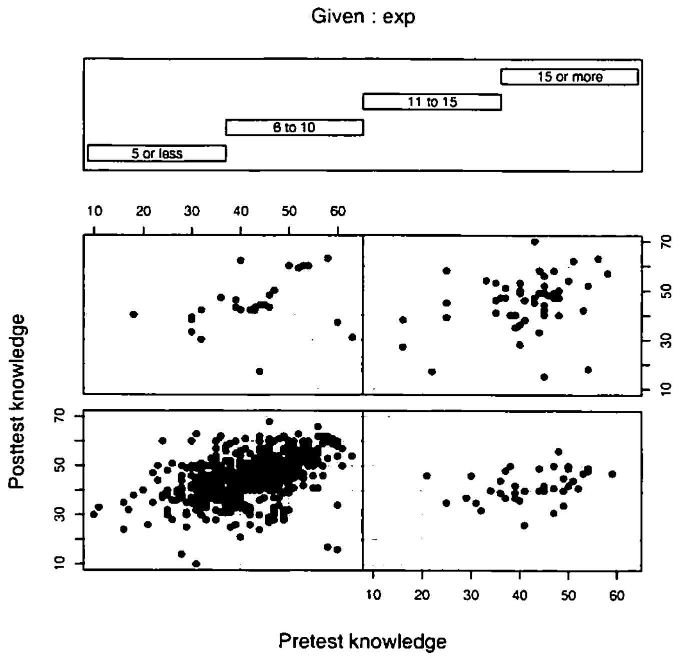


Figure 9.9: Knowledge by *experience*.

that suggests *experience* is related to posttest knowledge score (Table 9.23).

Table 9.23: Knowledge predicted by *experience*, combined levels.
ANOVA for model 7a (Experience combined)

	df	SS	F	Pr(>F)
Know T1	1	10921	175.27	<2e-16
Experience combined	1	98	1.58	0.21
Know T1*Experience combined	1	150	2.41	0.12
Resid	745	46423		

Model 8 : Centre

Figure 9.10 suggests some variation in the slopes of the centres, even though the results of the test in Table 9.24 indicate no strong evidence that posttest knowledge is related to *centre*. However, there is evidence that the interaction term is significant, which supports the earlier observation about the slopes. The strongest finding is that the posttest knowledge score seems to be more strongly dependent on the pretest score among participants from the centre of IDFR.

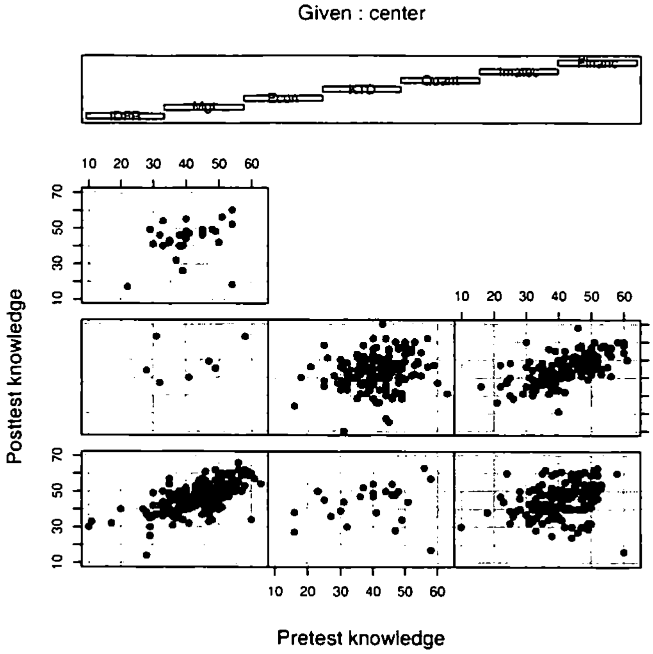


Figure 9.10: Knowledge by *centre.*

Model 9 : Course

The respondents for this study come from sixteen different courses, thus the *course* factor has sixteen levels. Figure 9.11 suggests that change in knowledge varies among the sixteen courses, but the majority of the boxplots indicates positive changes. The result of the analysis is as in Table 9.25. It appears that there is a slight evidence for an association between the posttest knowledge with *course*, after adjusting for the pretest score. If this is true, course number 12 appears to have a lower posttest knowledge score than the other courses. There is also evidence that course 12 has a different slope than the other courses. The coefficient of 0.31 suggests that the posttest knowledge score of this course is slightly more strongly dependent on the pretest knowledge score.

Model 10 : GHQ

The result of the ANCOVA test on the GHQ score is presented in Table 9.26. It shows no evidence at all of any association between the posttest knowledge and the score of the GHQ, after adjusting for the pretest score. In other words, the GHQ

Table 9.24: Knowledge predicted by *centre*.

ANOVA for model 8

	df	SS	F	Pr(>F)
Know T1	1	10921	181.86	<2.2e-16
Centre	6	861	2.39	0.03
Know T1*Centre	6	1671	4.64	0.0001
Resid	735	44139		

Coefficients:

	Est	*SE*	t	Pr(> $|t|$)
Intercept	29.55	2.34	12.64	<2e-16
Know T1	0.37	0.06	6.7	4.14e-11
Centre:IDFR	-10.22	3.28	-3.11	0.002
Centre:Mgt	6.35	5.14	1.24	0.22
Centre:Econ	3.79	3.56	1.06	0.29
Centre:KTD	5.96	10.42	0.57	0.57
Centre:Quant	4.42	3.61	1.22	0.22
Centre:Imatec	-8.64	3.46	-2.50	0.01
Know T1*IDFR	0.25	0.08	3.29	0.001
Know T1*Mgt	-0.17	0.12	-1.40	0.16
Know T1*Econ	-0.08	0.09	-0.97	0.33
Know T1*KTD	-0.05	0.25	-0.20	0.84
Know T1*Quant	-0.15	0.09	-1.78	0.08
Know T1*Imatec	0.19	0.08	2.39	0.02

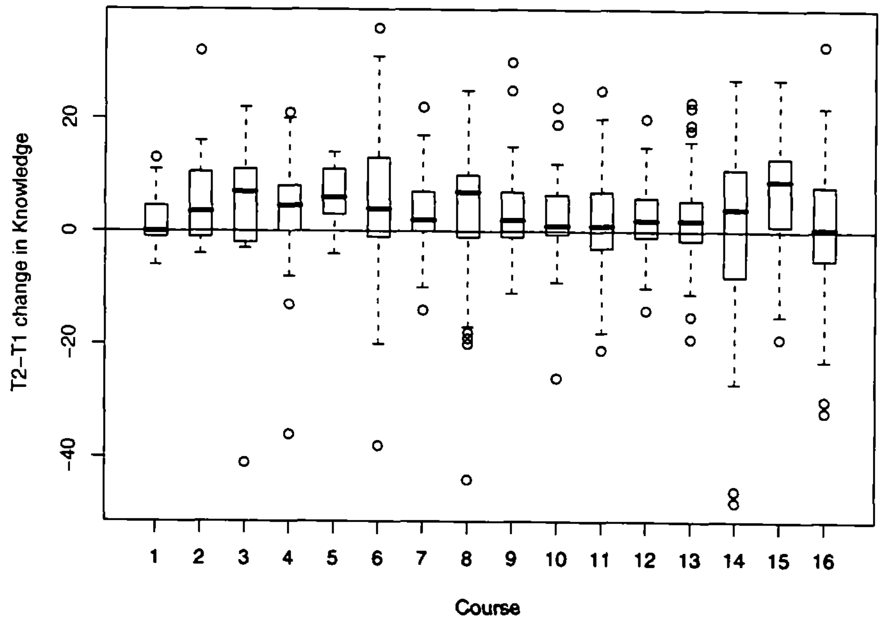


Figure 9.11: T2-T1 change in Knowledge by *course*

Table 9.25: Knowledge predicted by *course*.

ANOVA for model 9

	df	SS	F	Pr(>F)
Know T1	1	12259	179.12	<2.2e-16
Course	15	2144	2.09	0.009
Know T1*Course	15	2722	2.65	0.0006
Resid	728	49822		

Coefficients:

	Est	*SE*	t	Pr(> |*t*|)
Intercept	26.88	1.89	14.24	<2e-16
Know T1	0.42	0.05	9.16	<2e-16
Course:1	13.68	10.37	-1.32	0.19
Course:2	1.12	11.32	0.10	0.92
Course:3	9.88	7.97	1.24	0.22
Course:4	1.02	7.78	0.13	0.90
Course:5	-12.35	12.55	-0.94	0.33
Course:6	8.97	4.86	1.85	0.07
Course:7	-6.16	4.93	-1.25	0.21
Course:8	7.19	4.92	1.46	0.14
Course:9	-0.07	5.34	-0.01	0.99
Course:10	-5.76	5.44	-1.06	0.29
Course:11	-3.01	5.33	-0.57	0.57
Course:12	-12.52	4.48	-2.79	0.005
Course:13	-6.76	4.66	-1.45	0.15
Course:14	6.79	6.72	1.01	0.31
Course:15	10.44	7.84	1.33	0.18
Know T1*Course1	0.30	0.26	1.16	0.25
Know T1*Course2	0.05	0.28	0.18	0.86
Know T1*Course3	-0.36	0.20	-1.78	0.08
Know T1*Course4	-0.03	0.19	-0.14	0.89
Know T1*Course5	0.37	0.30	1.25	0.21
Know T1*Course6	-0.18	0.12	-1.52	0.13
Know T1*Course7	0.17	0.11	1.55	0.12
Know T1*Course8	-0.17	0.12	-1.43	0.15
Know T1*Course9	0.05	0.12	0.46	0.65
Know T1*Course10	0.17	0.12	1.41	0.16
Know T1*Course11	0.04	0.13	0.35	0.73
Know T1*Course12	0.31	0.10	3.07	0.002
Know T1*Course13	0.15	0.11	1.34	0.18
Know T1*Course14	-0.31	0.17	-1.85	0.06
Know T1*Course15	-0.19	0.19	-1.04	0.30

Table 9.26: Knowledge predicted by the GHQ. ANOVA for model 10

	df	SS	F	Pr(>F)
Know T1	1	12259	169.85	<2e-16
GHQ	1	0.40	0.006	0.94
Know T1*GHQ	1	124	1.71	0.19
Resid	756	54564		

score is not a predictor of the posttest knowledge, after the pretest knowledge is in the model.

9.4.5 Application Subscale

In this section, the pretest application score is the covariate, while the posttest application score is the dependent variable. After the basic model (Model 1), the demographic factors are included as predictors.

Model 1

The output for this basic model and its ANOVA result is presented in Table 9.27. The result suggests the pretest application is a significant predictor variable of the posttest score. The posttest score would increase by just 0.41 for every unit increase in the pretest score. The correlation coefficient between the variables is 0.401, giving an R^2 value of about 16.06%.

Table 9.27: Test result of Application basic model.

ANOVA for model 1

	df	SS	F	Pr(>F)
App T1	1	13105	142.97	2.2e-16
Resid	747	68474		

Coefficients:

	Est	*SE*	t	Pr(> $|t|$)
Intercept	29.89	1.78	16.77	<2e-16
App T1	0.41	0.03	11.96	<2e-16

Model 2 : Sex

Figure 9.12 shows no evidence to suggest significant differences between the two distributions. The test results in Table 9.28 suggests that when the pretest application score is adjusted for, *sex* is not a significant predictor variable of the posttest application.

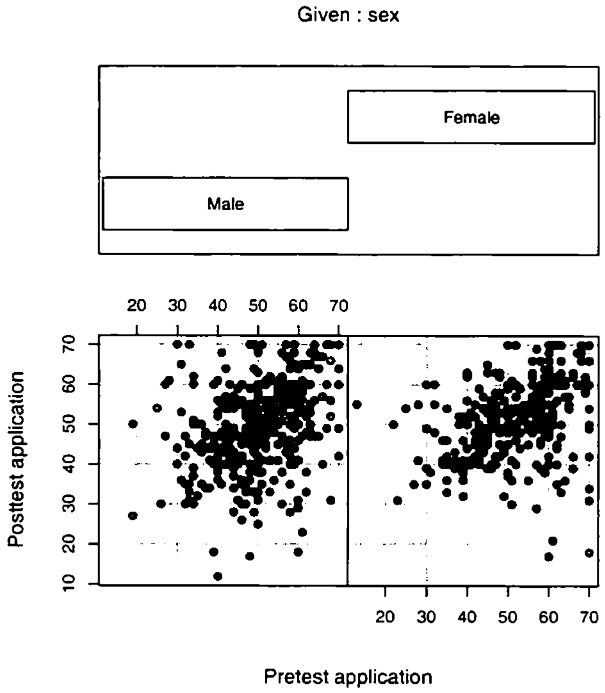


Figure 9.12: Application by *sex*.

Table 9.28: Application predicted by *sex*.
ANOVA for model 2

	df	SS	F	Pr(>F)
App T1	1	13201	143.93	<2e-16
sex	1	51	0.55	0.46
App T1*Sex	1	159	1.73	0.19
Resid	743	68150		

Table 9.29: Application predicted by Age Group.

ANOVA for model 3

	df	SS	F	Pr(>F)
App T1	1	13072	143.99	<2.2e-16
AgeGroup	5	180	0.40	0.85
App T1*Age	5	1459	3.21	0.007
Resid	734	66639		

Coefficients:

	Est	*SE*	t	Pr(> $\|t\|$)
Intercept	35.61	2.42	14.73	<2e-16
App T1	0.29	0.05	5.98	<3.6e-09
Age:<26	-11.28	4.06	-2.78	0.006
Age:26-30	-9.07	3.23	-2.81	0.005
Age:31-35	-2.87	4.29	-0.67	0.50
Age:36-40	2.36	6.54	0.36	0.72
Age:41-45	14.66	6.21	2.36	0.02
App T1*Age:<26	0.24	0.08	3.04	0.002
App T1*Age:26-30	0.18	0.06	2.78	0.006
App T1*Age:31-35	0.06	0.08	0.75	0.46
App T1*Age:36-40	-0.06	0.13	-0.43	0.67
App T1*Age:41-45	-0.30	0.13	-2.34	0.02

Model 3 : Age Group

Figure 9.13 suggests that there might be differences in the slopes among the age groups. Test results in Table 9.29 confirm this, and it involves the age groups of (i) <26 and (ii) 26 to 30. It appears that association between the posttest and the pretest application scores are slightly stronger for these groups. The implication is that for younger people there is a stronger positive dependency between pretest and posttest scores than for older people.

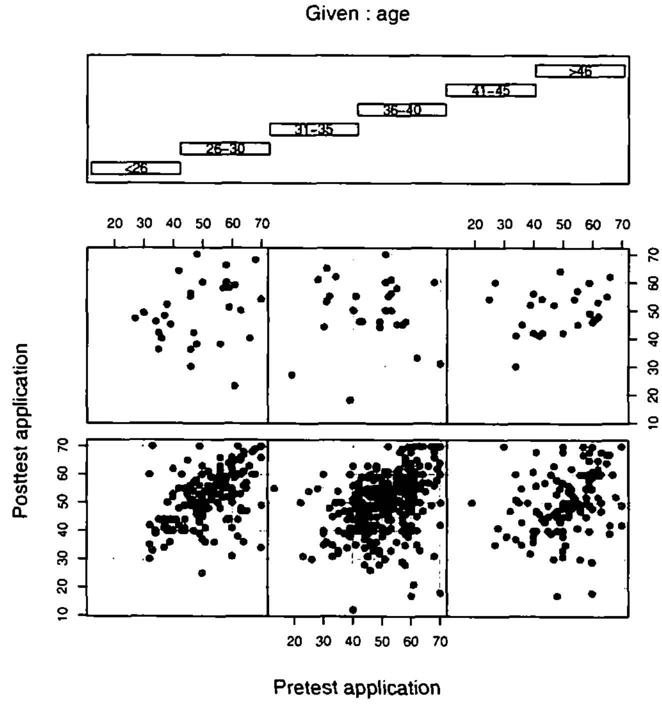


Figure 9.13: Application by *age group*.

Table 9.30: Application predicted by *ethnic group*. ANOVA for model 4

	df	SS	F	Pr(>F)
App T1	1	13078	142.79	<2e-16
Ethnic	3	402	1.46	0.22
App T1*Ethnic	3	197	0.72	0.54
Resid	739	67687		

Model 4 : Ethnic Group

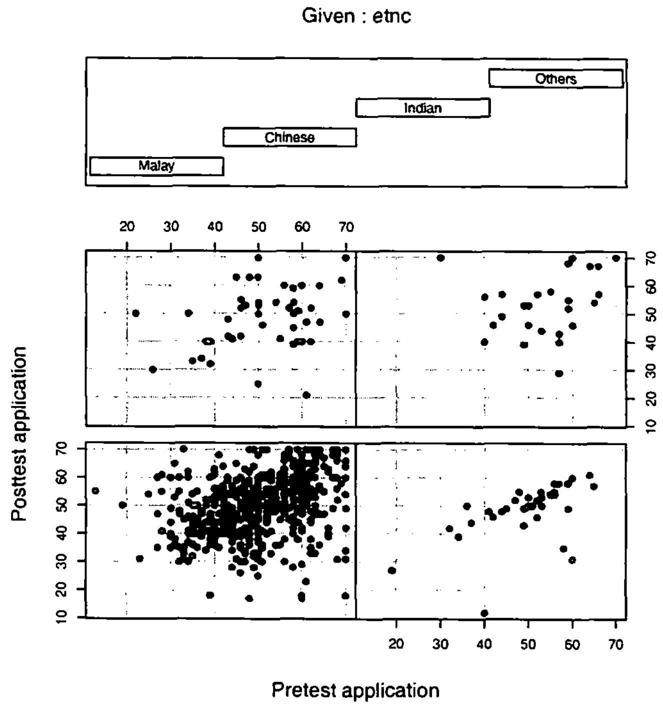


Figure 9.14: Application by *ethnic group.*

In Figure 9.14, the distributions of the ethnic groups seem to be similar. The ANOVA results (Table 9.30) suggest likewise, ie. ethnic group is not a significant predictor variable of the posttest application, when the pretest application score is adjusted for.

Table 9.31: Application predicted by *service sector.*

ANOVA for model 5

	df	SS	F	Pr(>F)
App Tl	1	13196	144.53	<2e-16
Service	2	149	0.82	0.44
App Tl*Service	2	26	0.14	0.87
Resid	740	67564		

Model 5 : Service Sector

Figure 9.15 does not suggest any differences among the three service sectors. Table 9.31 shows that where the participants work does not make any difference in the prediction of the posttest application score, when the pretest application score is

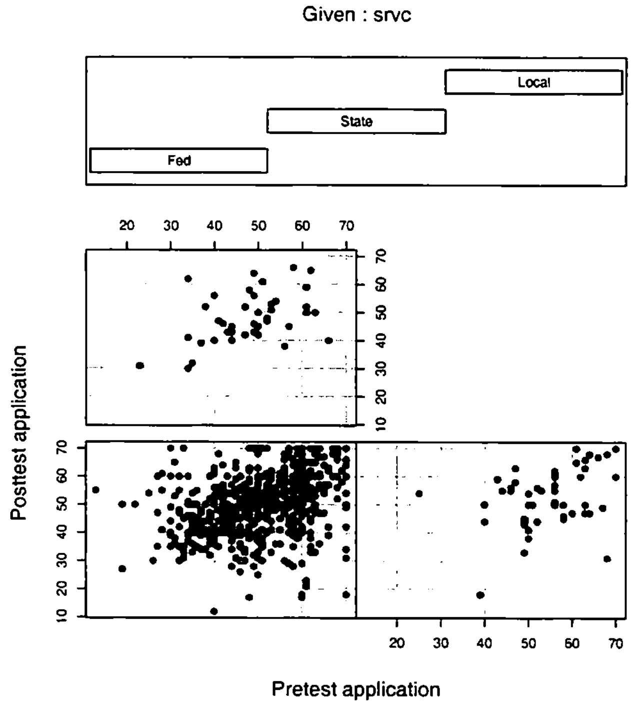


Figure 9.15: Application by *service sector.*

controlled for.

Model 6 : Service Group

Similar to the result for the *service sector* factor, the means of the posttest application

Table 9.32: Application predicted by *service group.*
ANOVA for model 6

	df	SS	F	Pr(>F)
App T1	1	12567	137.16	<2e-16
Group	1	254	2.77	0.10
App T1*Group	1	33	0.36	0.55
Resid	740	67802		

do not differ between the different service groups, as shown by the results in Table 9.32. *Service group* is not a significant predictor variable of the posttest application score.

Model 7 : Experience

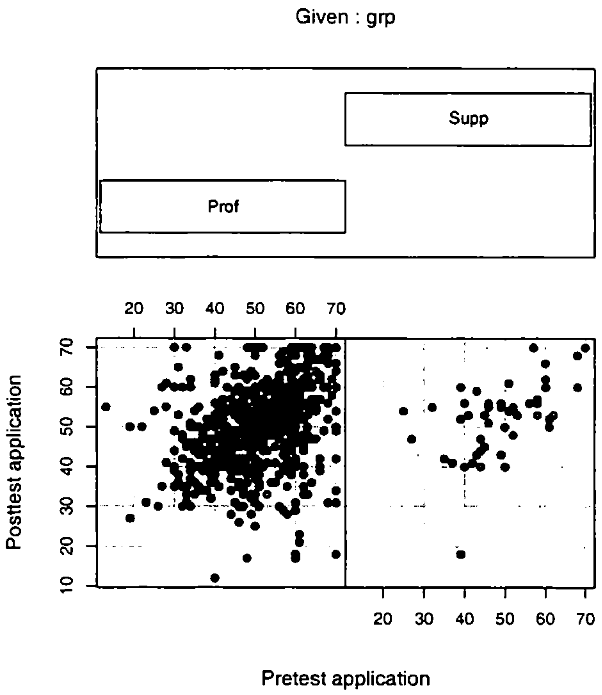


Figure 9.16: Application by *service group*.

Figure 9.17 suggests that there might be differences in the slopes and the intercepts of the different *experience* levels. Results of the ANOVA in Table 9.33 indicates a slight evidence that *experience* is related to the posttest application score when the pretest score is adjusted for. Participants with 5 years or less experience seem to have stronger pretest-posttest association than those with more experience.

Because of small sample sizes, the last three levels of *experience* are then combined and the analyses repeated. The results are presented in Table 9.34. With the experience levels combined, *experience* is not a significant predictor of the posttest application score, when the pretest application score is adjusted for. However, there is evidence that the slopes of the two experience groups are different. The association between the posttest and the pretest application scores is slightly stronger with the groups with 5 years or less experience. This finding is similar to that found with *age* (Model 3), where the association is stronger for younger participants.

Model 8 : Centre

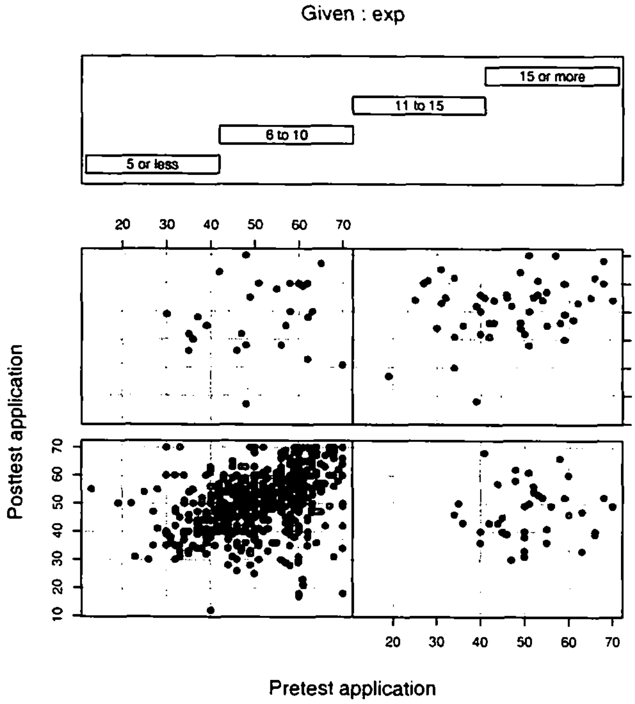


Figure 9.17: Application by *experience.*

Table 9.33: Application predicted by *experience.*

ANOVA for model 7

	df	SS	F	Pr(>F)
App T1	1	13105	145.99	<2e-16
Experience	3	1000	3.71	0.01
App T1*Experience	3	957	3.55	0.01
Resid	741	66517		

Coefficients:

	Est	*SE*	t	Pr(> $\|t\|$)
Intercept	37.28	3.41	10.92	<2e-16
App T1	0.24	0.07	3.64	0.0003
Exp:5 or less	-10.11	3.69	-2.74	0.006
Exp:6 to 10	7.36	7.07	1.04	0.30
Exp:11 to 15	-0.50	7.12	-0.07	0.94
App T1*Exp:5 or less	0.22	0.07	3.11	0.002
App T1*Exp:6 to 10	-0.20	0.14	-1.45	0.15
App T1*Exp:10 to 15	-0.01	0.14	-0.08	0.94

Table 9.34: Application predicted by *experience*, combined levels.

ANOVA for model 7a (Experience combined)

	df	SS	F	Pr(>F)
App T1	1	13105	145.29	<2.2e-16
Experience combined	1	97	1.08	0.30
App T1*Experience combined	1	1177	13.05	0.0003
Resid	745	67200		

Coefficients:

	Est	*SE*	t	Pr(> $\|t\|$)
Intercept	34.52	2.22	15.53	<2e-16
App T1	0.30	0.04	6.99	<6e-12
exp:5 or less	-7.35	2.22	-3.31	0.001
App T1*exp:5 or less	0.16	0.04	3.61	0.0003

Figure 9.18 suggests that there are variations in the slopes and intercepts of the regression lines of the different centres. Looking at the results in Table 9.35, there is an evidence that *centre* is a significant factor in predicting the posttest application score when the pretest application score is controlled for. Both the centres of IDFR and the centre of IMATEC indicate lower posttest application scores. The result of the interaction term supports the suggestion of the plots. For the same two centres, the relationship between the posttest and the pretest application scores appears to be stronger. This is also similar in part to the relationship found between knowledge and *centre* (Model 8 on page 220).

Model 9 : Course

Figure 9.19 shows the distribution of the T2-T1 change in application by *course*. There appears to be some variation in the change among the courses, with some courses showing positive changes, while some others showing negative changes.

The result of the ANCOVA is presented in Table 9.36. It suggests that the posttest application is associated with *course* after adjusting for the pretest score. This means that changes in application are likely to differ among the different *courses*. Course 11 seems to have a slightly lower posttest application score than the other courses, and slightly stronger dependence on pretest score.

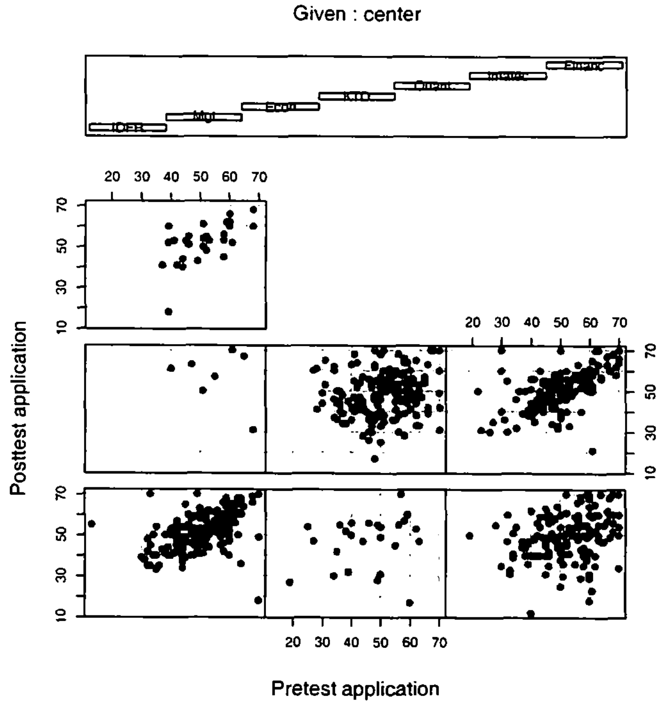


Figure 9.18: Application by *centre.*

Model 10 : GHQ

The test on the GHQ as a possible predictor of the posttest application suggests only a slight evidence, as presented in Table 9.37. It is not very likely that the GHQ score is a predictor of the posttest application, after adjusting for the pretest score.

9.4.6 Importance Subscale

The factor of importance measures the participants' personal view on the importance of learning and enhancing their knowledge in the subject areas. The posttest importance score is the dependent variable, and the pretest importance score is the covariate.

Model 1

The output for this model and its ANOVA result is presented in Table 9.38. The

Table 9.35: Application predicted by *centre*.

ANOVA for model 8

	df	SS	F	Pr(>F)
App T1	1	13105	152.61	<2.2e-16
Centre	6	2185	4.24	0.0003
App T1*Centre	6	3171	6.15	2.6e-06
Resid	735	63118		

Coefficients:

	Est	*SE*	t	Pr(> $\|t\|$)
Intercept	36.29	3.65	9.94	<2e-16
App T1	0.30	0.07	4.42	<1.2e-05
Centre:IDFR	-12.93	4.66	-2.77	0.006
Centre:Mgt	3.43	6.87	0.50	0.62
Centre:Econ	-4.88	4.90	-1.00	0.32
Centre:KTD	42.48	18.16	2.34	0.02
Centre:Quant	5.60	4.79	1.17	0.24
Centre:Imatec	-14.44	4.81	-3.00	0.003
App T1*Centre:IDFR	0.27	0.09	3.15	0.002
App T1*Centre:Mgt	-0.13	0.14	-0.95	0.34
App T1*Centre:Econ	0.05	0.09	0.54	0.59
App T1*Centre:KTD	-0.69	0.32	-2.13	0.03
App T1*Centre:Quant	-0.16	0.09	-1.82	0.07
App T1*Centre:Imatec	0.28	0.09	3.09	0.002

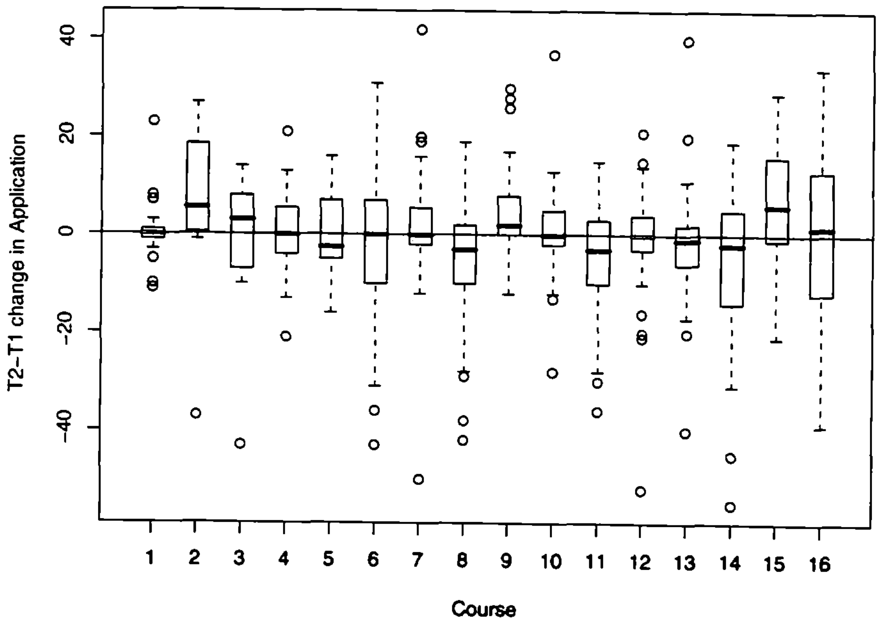


Figure 9.19: T2-T1 change in Application by *course*

Table 9.36: Application predicted by *course*.

ANOVA for model 9

	df	SS	F	Pr(>F)
App T1	1	14940	156.84	<2.2e-16
Course	15	5314	3.72	2.4e-06
App T1*Course	15	5756	4.03	4.4e-07
Resid	728	69348		

Coefficients:

	Est	*SE*	t	Pr(> $\|t\|$)
Intercept	29.28	2.26	12.94	<2e-16
App T1	0.40	0.05	8.90	<2e-16
Course:1	-14.51	9.44	-1.54	0.12
Course:2	6.13	8.66	0.71	0.45
Course:3	6.04	11.80	0.51	0.61
Course:4	-12.23	10.49	-1.17	0.24
Course:5	0.08	14.89	0.005	0.99
Course:6	2.29	5.66	0.40	0.69
Course:7	-7.25	5.51	-1.32	0.19
Course:8	-1.53	6.46	-0.24	0.81
Course:9	2.42	6.20	0.39	0.70
Course:10	-6.84	6.97	-0.98	0.33
Course:11	-15.45	6.53	-2.37	0.02
Course:12	-9.94	6.21	-1.60	0.11
Course:13	-11.96	6.08	-1.97	0.05
Course:14	16.63	13.35	1.25	0.21
Course:15	17.69	8.83	2.00	0.05
App T1*Course1	0.28	0.20	1.38	0.17
App T1*Course2	-0.03	0.17	-0.19	0.85
App T1*Course3	-0.31	0.27	-1.13	0.26
App T1*Course4	0.28	0.20	1.40	0.16
App T1*Course5	0.04	0.28	0.14	0.88
App T1*Course6	-0.04	0.11	-0.42	0.68
App T1*Course7	0.20	0.10	1.86	0.06
App T1*Course8	-0.02	0.12	-0.14	0.88
App T1*Course9	0.05	0.12	0.37	0.71
App T1*Course10	0.20	0.13	1.53	0.13
App T1*Course11	0.25	0.12	2.06	0.04
App T1*Course12	0.22	0.12	1.80	0.07
App T1*Course13	0.23	0.12	2.02	0.04
App T1*Course14	-0.44	0.26	-1.70	0.09
App T1*Course15	-0.31	0.18	-1.69	0.09

Table 9.37: Application predicted by the GHQ.

ANOVA for model 10

	df	SS	F	Pr(>F)
App T1	1	14940	141.42	2e-16
GHQ	1	539	5.10	0.02
App T1*GHQ	1	9	0.09	0.77
Resid	756	79870		

Table 9.38: Test result of Importance model 1.

ANOVA for model 1

	df	SS	F	Pr(>F)
Imp T1	1	15459	166.47	<2.2e-16
Resid	747	69369		

Coefficients:

	Est	*SE*	t	Pr(> $|t|$)
Intercept	29.28	2.01	14.58	<2e-16
Imp T1	0.45	0.04	12.90	<2e-16

results suggest that the pretest importance is a significant predictor variable of the posttest importance score. The correlation coefficient between them is 0.427, suggesting the percentage of explained variation of about 18.22%. The value of the intercept is almost 30, which is the estimated value of the posttest importance without a pretest input. It is estimated that one unit increase in the pretest importance score would result in 0.45 unit increase in the posttest score.

Model 2 : Sex

Both of the distributions in Figure 9.20 appear to be similar. The test results for this model are in Table 9.39. There is no evidence to suggest that *sex* is a significant predictor variable for the posttest importance score. The average score of the posttest importance does not differ between male and female participants, after adjusting for the pretest score. However, Figure 9.20 also indicates many high scores, which casts doubt on the validity of the analysis.

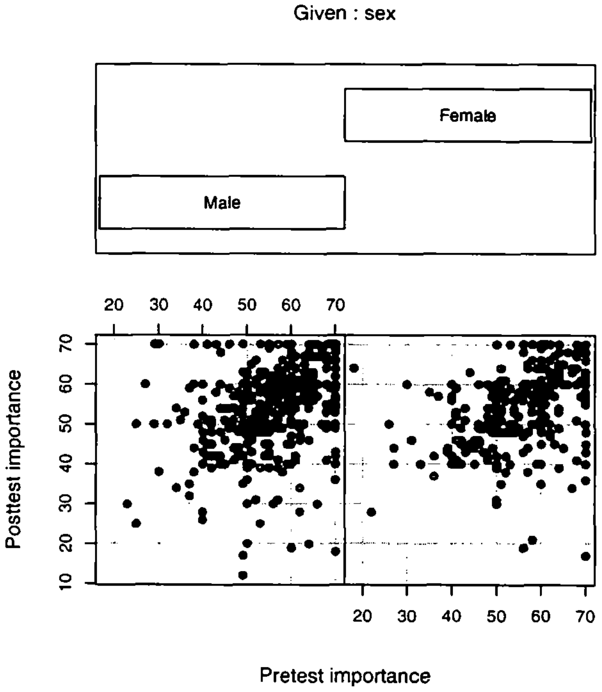


Figure 9.20: Importance by *sex*.

Table 9.39: Importance predicted by *sex*.
ANOVA for model 2

	df	SS	F	Pr(>F)
Imp T1	1	15448	165.76	<2e-16
sex	1	7	0.07	0.79
Imp T1*Sex	1	104	1.12	0.29
Resid	743	69241		

Model 3 : Age Group

Figure 9.21 does not indicate any evidence of differences among the age groups.

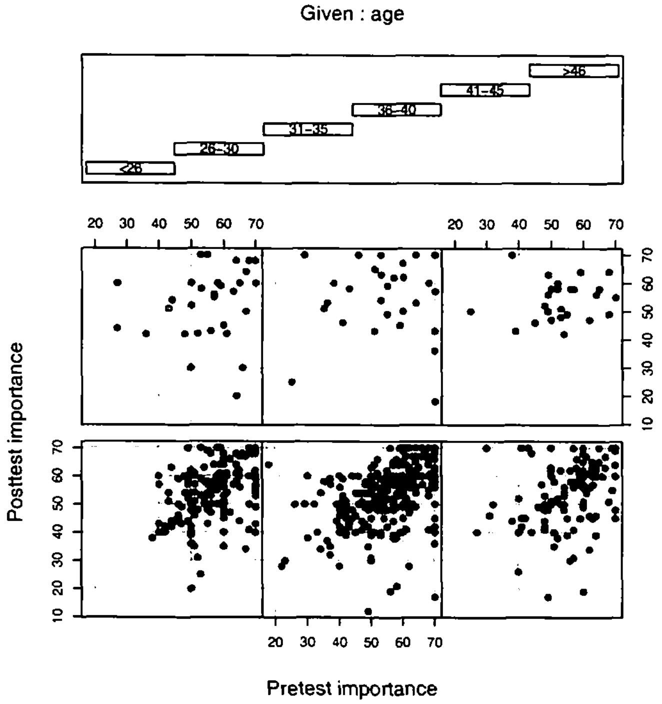


Figure 9.21: Importance by *age group*.

The coefficients and ANOVA output for this model is as in the Table 9.40. Age group does not seem to be a significant factor in predicting the posttest importance score when the pretest importance score is controlled for.

However, it seems likely some age groups have different slopes, as suggested by the result of F test on the interaction term. Looking at the coefficient table, two age groups seem to be just that, namely the age groups of (i) <26 years, and (ii) 26 to 30 years. It is suggested that the relationship between the posttest and the

Table 9.40: Importance predicted by *age group*.

ANOVA for model 3

	df	SS	F	Pr(>F)
Imp T1	1	15419	168.52	<2.2e-16
AgeGroup	5	141	0.30	0.91
Imp T1*AgeGroup	5	2048	4.48	0.0005
Resid	734	67158		

Coefficients:

	Est	*SE*	t	Pr(> $\|t\|$)
Intercept	37.70	2.90	13.01	<2e-16
Imp T1	0.30	0.05	5.66	<2.17e-08
age:<26	-14.65	4.89	-3.00	0.003
age:26-30	-12.71	3.68	-3.46	0.0006
age:31-35	-7.00	5.13	-1.36	0.17
age:36-40	5.02	7.74	0.65	0.52
age:41-45	17.96	7.00	2.57	0.01
Imp T1*Age:<26	0.27	0.09	3.14	0.002
Imp T1*Age:26-30	0.24	0.07	3.61	0.0003
Imp T1*Age:31-35	0.13	0.09	1.40	0.16
Imp T1*Age:36-40	-0.12	0.14	-0.84	0.40
Imp T1*Age:41-45	-0.31	0.13	-2.49	0.01

pretest importance scores is stronger in these groups. This finding shows similar relationship as for application (Page 229).

Model 4 : Ethnic Group

Figure 9.22 does not suggest any differences among the distributions of the different ethnic groups. The results in Table 9.41 indicate no evidence that *ethnic group* predicts the posttest importance. The average posttest importance score does not differ among the different ethnic groups, after adjusting for the pretest score.

Table 9.41: Importance predicted by *ethnic group*.

ANOVA for model 4

	df	SS	F	Pr(>F)
Imp T1	1	15414	166.98	<2e-16
Ethnic	3	605	2.19	0.09
Imp T1*Ethnic	3	521	1.88	0.13
Resid	739	68215		

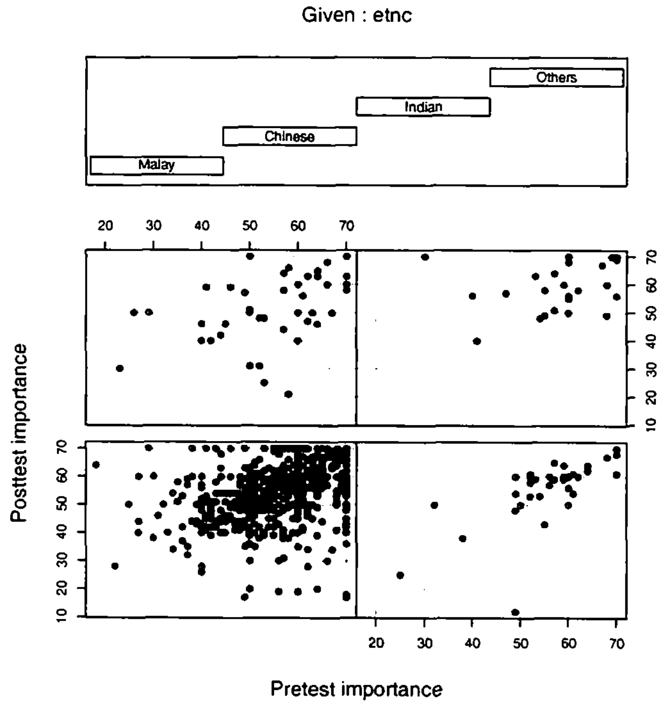


Figure 9.22: Importance by *ethnic group*.

Model 5 : Service Sector

There is no indication of differences indicated by Figure 9.23. The results in Table 9.42 also suggest that *service sector* is not a significant predictor variable of the posttest importance score, when the pretest importance is adjusted for.

Table 9.42: Importance predicted by *service sector*.
ANOVA for model 5

	df	SS	F	Pr(>F)
Imp T1	1	15359	164.48	<2e-16
Service	2	186	1.00	0.37
Imp T1*Service	2	28	0.15	0.86
Resid	740	69097		

Model 6 : Service Group

The results for *service group* is in Table 9.43. *Service group* does not appear to be a significant predictor variable of the posttest importance score when the pretest

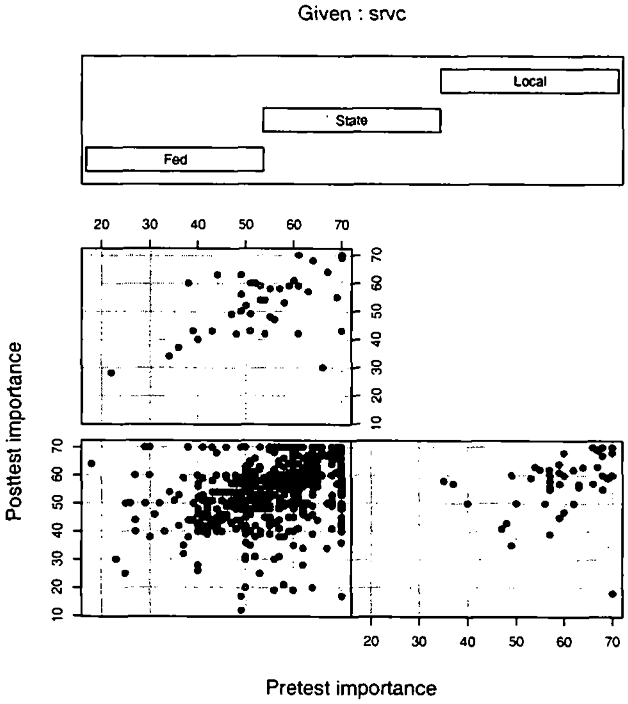


Figure 9.23: Importance by *service sector.*

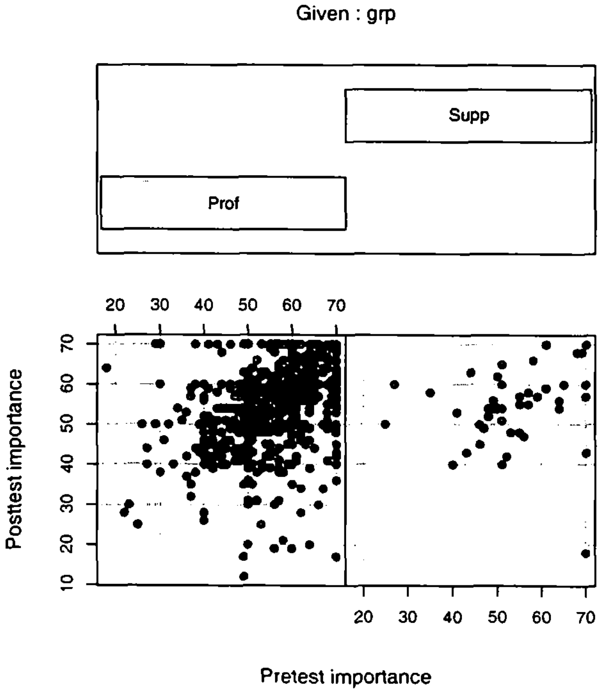


Figure 9.24: Importance by *service group.*

Table 9.43: Importance predicted by *service group*. ANOVA for model 6

	df	SS	F	Pr(>F)
Imp T1	1	14946	160.72	<2e-16
Group	1	81	0.87	0.35
Imp T1*Group	1	260	2.79	0.10
Resid	740	68817		

importance score is adjusted for. This finding agrees with the plots of pretest-posttest importance by service group in Figure 9.24.

Model 7 : Experience

Figure 9.25 does not seem to suggest any differences among the four levels of experience. The results of ANOVA in Table 9.44 suggest a slight evidence that *experience* is related to the posttest importance score, when the pretest importance score is adjusted for. There is also a slight evidence regarding the interaction.

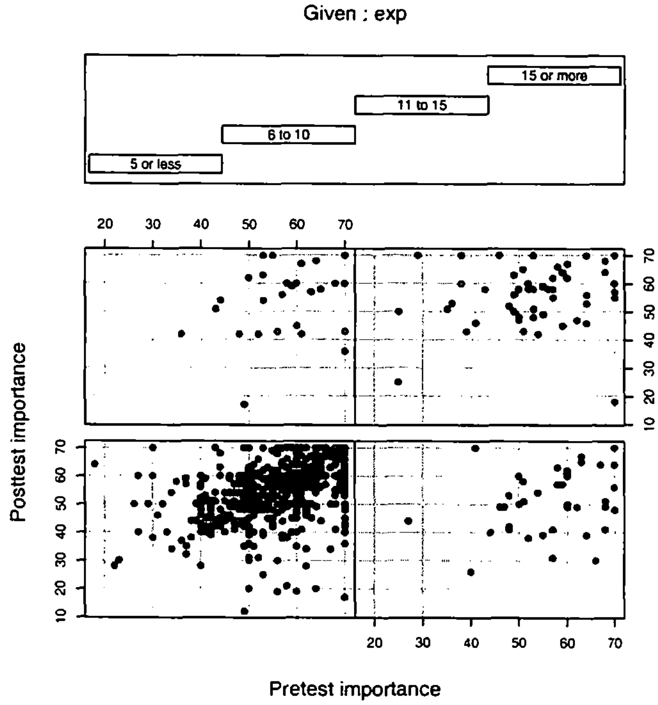


Figure 9.25: Importance by *experience*.

In the next analysis, the last three levels of *experience* are combined because

Table 9.44: Importance predicted by *experience*.

ANOVA for model 7

	df	SS	F	Pr(>F)
Imp T1	1	15459	169.71	<2e-16
Experience	3	905	3.31	0.02
Imp T1*Experience	3	966	3.53	0.01
Resid	741	67499		

Coefficients:

	Est	*SE*	t	Pr(> $\|t\|$)
Intercept	35.67	4.14	8.61	<2e-16
Imp T1	0.32	0.07	4.45	10.0e-06
exp:5 or less	-9.26	4.42	-2.09	0.04
exp:6 to 10	-1.53	7.61	-0.20	0.84
exp:11 to 15	-1.93	9.53	-0.20	0.84
Imp T1*exp:5 or less	0.19	0.08	2.40	0.02
Imp T1*exp:6 to 10	-0.02	0.13	-0.19	0.85
Imp T1*exp:11 to 15	0.03	0.16	0.18	0.85

of small sample sizes, creating just two levels, namely: (i) 5 years or less, and (ii) more than 5 years. The results of analyses are as in Table 9.45. There is now a strong evidence for the interaction term, suggesting that the slopes of the two experience groups are different. The relationship between the posttest and the prestest importance scores is stronger among the group with 5 years or less experience. This is similar to the results for application on page 231.

Model 8 : Centre

Figure 9.26 indicates some variations in the slopes of the pretest-posttest distributions of the different centres. However, ANOVA results in Table 9.46 suggest no evidence supporting that indication, ie. the factor of *centre* is a significant predictor variable of the posttest importance score when the pretest importance score is adjusted for.

Model 9 : Course

The distribution of the T2-T1 change in importance by *course* is presented in Figure 9.27. The boxplots suggest some variation in the change among the *courses*, but

Table 9.45: Importance predicted by *experience*, combined levels.

ANOVA for model 7a (Experience combined)

	df	SS	F	Pr(>F)
Imp T1	1	15459	168.90	<2.2e-16
Experience combined	1	85	0.93	0.34
Imp T1*Experience combined	1	1097	11.99	0.0006
Resid	745	68187		

Coefficients:

	Est	*SE*	t	Pr(> $\|t\|$)
Intercept	34.87	2.62	13.30	<2e-16
Imp T1	0.35	0.05	7.45	<2.6e-13
exp. combined:5 or less	-8.46	2.62	-3.23	0.001
Imp T1*Exp. combined:5 or less	0.16	0.05	3.46	0.0006

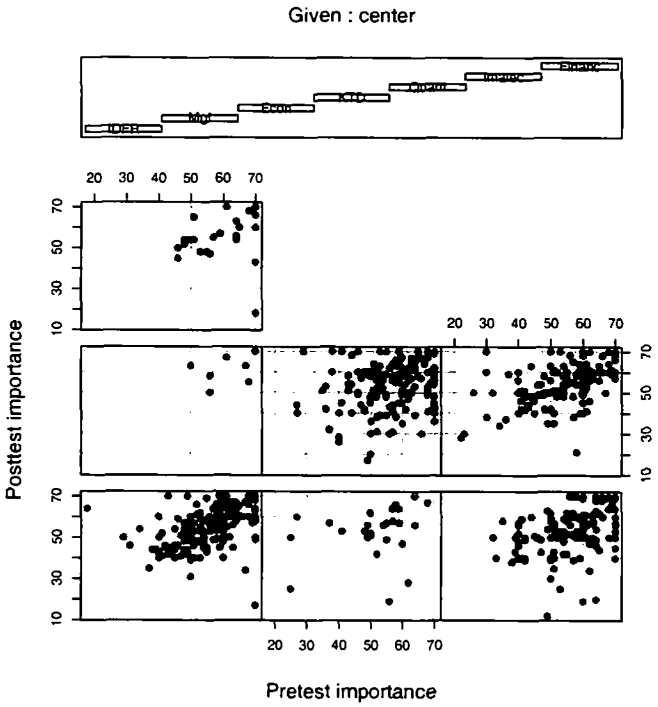


Figure 9.26: Importance by *centre*.

Table 9.46: Importance predicted by *centre*.

ANOVA for model 8

	df	SS	F	Pr(>F)
Imp T1	1	15459	170.08	<2e-16
Centre	6	1076	1.97	0.07
Imp T1*Centre	6	1488	2.73	0.01
Resid	735	66805		

in most courses the change does not seem to be far from zero.

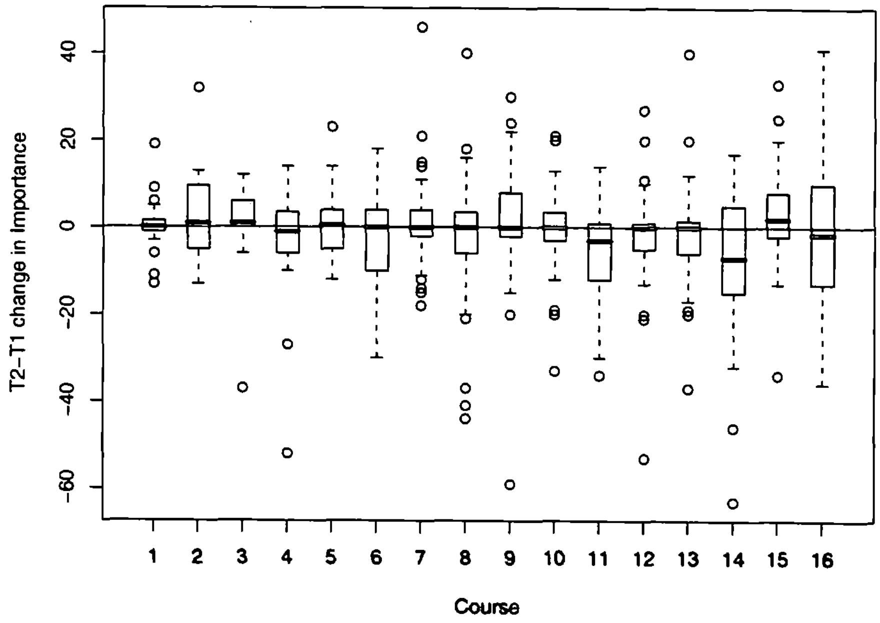


Figure 9.27: T2-T1 change in Importance by *course*

The result of the test as in Table 9.47 suggests with some evidence that *course* is associated with the posttest importance. The interaction between pretest importance and *course* is also significant. The coefficient implies that for this course there is a strong negative association between the posttest and the pretest importance scores.

Model 10 : GHQ

The result of this model is in Table 9.48. It does not indicate any evidence for an association between the GHQ score and the posttest importance, after the pretest score is controlled.

9.4.7 Overall Summary

All the results of ANCOVA analysis are presented in Table 9.49 on page 251. Probability values (P-values) are shown for the factor term and the interaction term of each of the factors. Probability values which are less than 0.01 are highlighted, suggesting evidence for significance. The results suggest that the posttest knowledge score is generally not predicted by the seven demographic variables after adjusting for the pretest knowledge score, except for the factor of *course*, where the average posttest score might differ among the different *courses.*

Two demographic factors appear to be predictors of the posttest application score when the pretest scores are adjusted for. The factors are *centre* and *course*. This seems to suggest that the means of the posttest application score differ among the different centres, or among the different courses. For the importance subscale, there is an evidence that the average posttest scores differ among the different courses, when the pretest importance score is controlled for. The posttest score does not differ between or among the levels of any other demographic factors.

For interaction terms, there are more cases with evidence that supports significance. The interactions between *age* and pretest application, as well as *age* and pretest importance seem to be significant. This suggest that the slopes among the different age groups are different when tested with the two scales. The same applies to the factor of *combined experience*, which has only two levels. The slopes between the two levels appear to be different when tested with application and importance.

There are also evidences for significant interaction terms involving the factor of *centre*, when tested with knowledge and application. The factor of *course* meanwhile, appear to have its interaction terms significant when tested with all three scales; knowledge, application, and importance. This suggests that the slopes of the relationship between the posttest and the prestest scores of all three scales are different among the different courses.

Table 9.47: Importance predicted by *course*.

ANOVA for model 9

	df	SS	F	Pr(>F)
Imp T1	1	16369	168.80	<2.2e-16
Course	15	3297	2.27	0.004
Imp T1*Course	15	4390	3.02	9.9e-05
Resid	728	70595		

Coefficients:

	Est	*SE*	t	Pr(> $\lvert t \rvert$)
Intercept	28.53	2.60	10.98	<2e-16
Imp T1	0.46	0.05	9.96	<2e-16
Course:1	-16.85	9.25	-1.82	0.07
Course:2	4.36	9.20	0.47	0.64
Course:3	-18.67	14.93	-1.25	0.21
Course:4	5.03	12.35	0.41	0.68
Course:5	0.57	14.15	0.04	0.97
Course:6	-1.94	7.46	-0.26	0.79
Course:7	-0.36	6.68	-0.05	0.96
Course:8	0.19	6.03	0.03	0.97
Course:9	9.23	6.75	1.37	0.17
Course:10	-5.55	7.40	-0.75	0.45
Course:11	-10.85	7.20	-1.51	0.13
Course:12	-12.91	6.44	-2.01	0.05
Course:13	-7.38	6.86	-1.08	0.28
Course:14	-1.87	18.70	-0.10	0.92
Course:15	26.37	10.47	2.52	0.01
Imp T1*Course1	0.30	0.19	1.60	0.11
Imp T1*Course2	-0.01	0.16	-0.05	0.96
Imp T1*Course3	0.31	0.29	1.05	0.30
Imp T1*Course4	-0.08	0.20	-0.38	0.70
Imp T1*Course5	0.04	0.25	0.18	0.86
Imp T1*Course6	0.03	0.13	0.23	0.82
Imp T1*Course7	0.05	0.12	0.43	0.67
Imp T1*Course8	-0.02	0.11	-0.12	0.85
Imp T1*Course9	-0.13	0.12	-1.04	0.30
Imp T1*Course10	0.13	0.13	1.01	0.31
Imp T1*Course11	0.14	0.12	1.17	0.24
Imp T1*Course12	0.22	0.12	1.93	0.05
Imp T1*Course13	0.13	0.12	1.09	0.28
Imp T1*Course14	-0.09	0.32	-0.28	0.78
Imp T1*Course15	-0.46	0.20	-2.35	0.02

Table 9.48: Importance predicted by the GHQ. ANOVA for model 10

	df	SS	F	Pr(>F)
Imp T1	1	16369	158.67	<2e-16
GHQ	1	275	2.66	0.10
Imp T1*GHQ	1	18	0.18	0.97
Resid	756	77988		

Table 9.49: Summary of the ANCOVA on the Learning subscales.

		P-values of ANOVA		
Factor	**Term**	*Knowledge*	*Application*	*Importance*
Sex	Factor	0.71	0.46	0.79
	Interaction	0.69	0.19	0.29
Age	Factor	0.24	0.85	0.91
	Interaction	0.04	**0.007**	**0.0005**
Ethnic	Factor	0.19	0.22	0.09
	Interaction	0.55	0.54	0.13
Service sector	Factor	0.07	0.44	0.37
	Interaction	0.86	0.87	0.86
Service group	Factor	0.68	0.10	0.35
	Interaction	0.50	0.55	0.10
Experience	Factor	0.07	0.01	0.02
	Interaction	0.40	0.01	0.01
Experience combined	Factor	0.21	0.30	0.34
	Interaction	0.12	**0.0003**	**0.0006**
Centre	Factor	0.03	**0.0003**	0.07
	Interaction	**0.0001**	**2.6e-06**	0.01
Course	Factor	**0.009**	**2.4e-06**	**0.004**
	Interaction	**0.0006**	**4.5e-07**	**9.9e-05**
GHQ	Factor	0.94	0.02	0.10
	Interaction	0.19	0.77	0.67

Interaction between demographic factors

The results of the ANCOVA as presented in Table 9.49 in the previous section suggest that posttest scores of knowledge and importance are dependent on *course*, while the posttest score of application is dependent on both *centre* and *course.* In cases where two or more demographic variables seem to be related to the posttest scores, it is also of interest to examine the interactions. In this case, there is only one such instance, ie. posttest application and the interaction of *centre* and *course.*

However, both *centre* and *course* are hierarchically related. There are a total of 16 different courses and there are 7 centres, but each course is uniquely under a specific centre. The courses and the centre they relate to are presented in the following table:

Courses under the Centres.

Courses	**Centres**
4 and 16	1 (Management)
6, 7 and 9	2 (Economic)
3	3 (KTD)
12, 15 and 17	4 (Quantitative)
1, 10 and 14	5 (Imatec)
5	6 (Finance)
8, 11 and 13	7 (IDFR)

A sophisticated method to analyze such data is to take account explicitly of the hierarchical nature of the data. However, because of time constraint, we employ a simpler approach. To gauge whether *course* is needed in addition to *centre*, we fit Model 1 below. The ANOVA in Table 9.50 suggests that *centre* is insufficient, and that *course* must be included in the model.

Model 1:

Posttest = Pretest + *Centre* + *Course* + Pretest**Centre* + Pretest**Course*

Since *centre* and *course* are hierarchically related, an interaction term between them would make no sense, thus it is not included in the model. In Model 1, the *centre* term is included first before *course.* As expected, all four terms, ie. the two factors (*centre* and *course*) and the two interactions, indicate evidence of significance. Because of the hierarchical nature of these terms, there are reduced degrees of freedom for the *course* terms. In addition, the inferences for the addition of the *course* term are somewhat misleading because these ought to be treated as nested within *centre*.

Table 9.50: ANOVA results of Model 1 with demographic interaction.

ANOVA	Df	SS	F value	Pr(>F)
App T1	1	14940	156.84	<2.2e-16
Centre	6	2596	4.54	0.00016
Course	9	2718	3.17	0.0009
App T1*Centre	6	2860	5.00	4.9e-05
App T1*Course	9	2895	3.38	0.0005
Residuals	728	69348		

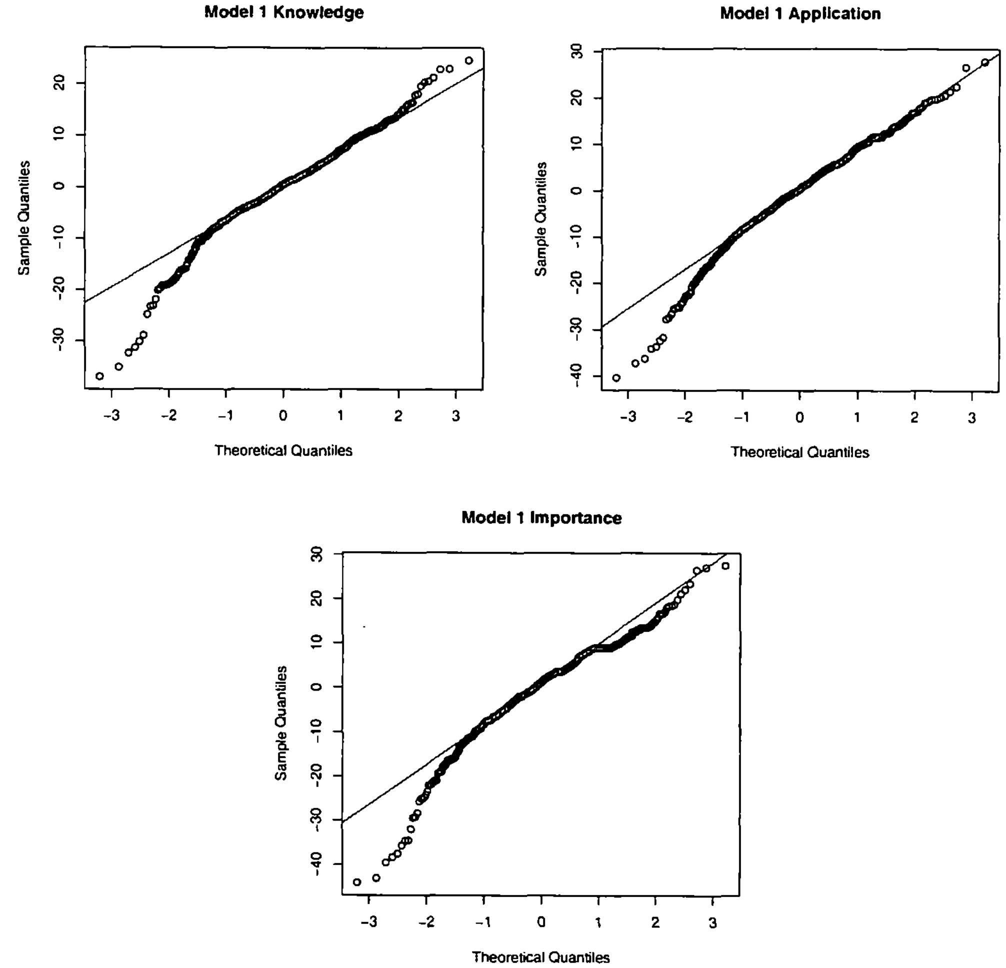
With more sophisticated analysis, we may be able to conclude that *centre* and *course* have separate effects on posttest application score. With this analysis however, the simple message is that there is a *course* effect beyond the *centre* effect. If we wish to fit just one variable, we fit *course*.

Residuals

Figure 9.28 shows the normal QQ plots of the residuals of the three basic models, namely the basic models of knowledge, application, and importance. In each basic model, the posttest score is the dependent and the pretest score is the only explanatory variable. The plots indicate that the residuals deviate from Normality, especially at the lower ends of the distributions.

Figures 9.29, 9.30 and 9.31 show the normal QQ plots of the residuals of the seven models which have some evidence of significance. We can see generally that the plots look very similar to those in Figure 9.28. Overall, the residuals of the model fits are not perfectly Normally distributed. We hope that this does not affect the analysis too much. The fact that the residual plots are similar among the models suggests that the outliers are not dependent on any of the demographic factors relevant to the models.

In each of the three Learning factors, the correlation between the dependent variable (posttest score) and the covariate (pretest score) is not very high. For the



Model 1 Knowledge

Model 1 Application

Model 1 Importance

Figure 9.28: Normality plots of basic models.

scale of knowledge, the correlation coefficient between the pretest and posttest scores is 0.432. For application the correlation is 0.400, while for importance it is 0.430. Thus R-squared figures are 18.69%, 16.04% and 18.48% for knowledge, application, and importance respectively. These values imply low percentages of variance of the posttest data explained by the pretest data. On average, however, the relationship seems to be very strong, but there is a lot of individual variation.

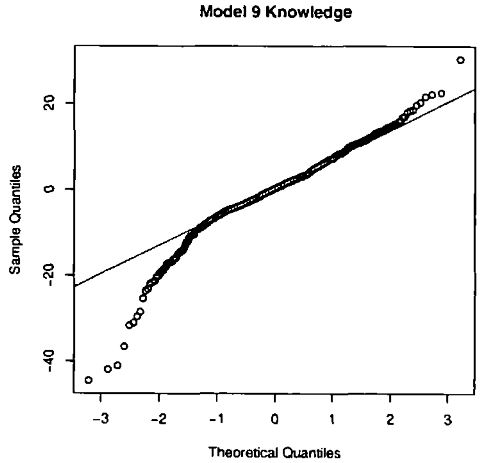


Figure 9.29: Residual plot of significant model for Knowledge.

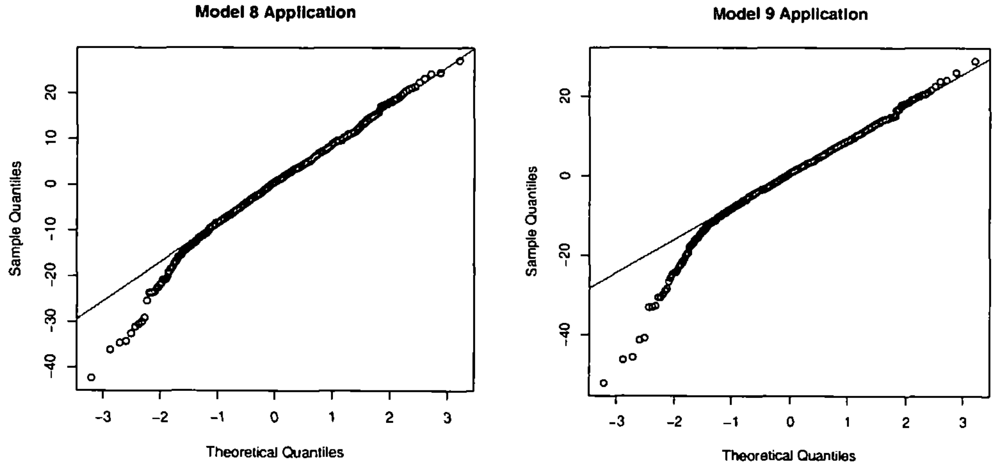


Figure 9.30: Residual plots of significant models for Application.

9.4.8 Relationship between GHQ and CEQ

It is of interest to know whether the general mental health of the participants has any association with their experience of the training. We can examine this by looking at the relationship between the GHQ scores and the scores of the CEQ subscales. For this purpose, the GHQ scoring (0, 0, 1, 1) is used, instead of 1, 2, 3 and 4 as used previously in this thesis. The number of 1's is then counted for each participant, and if the number exceeds 3, he or she is categorized as Stressed. Otherwise he or she is categorized as Not Stressed. Table 9.51 on page 256 indicates the number

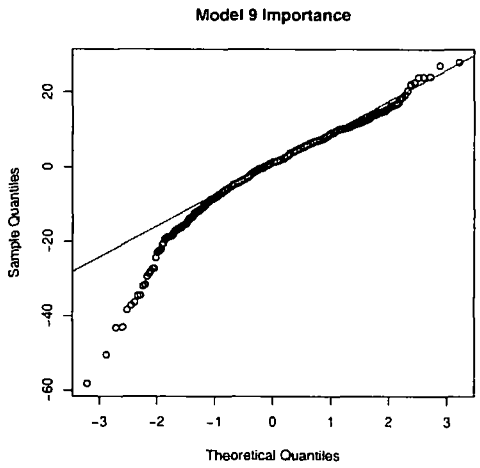


Figure 9.31: Residual plot of significant model for Importance.

of samples in each GHQ category, as well as the mean scores of each of the CEQ subscales.

Table 9.51: Means of CEQ by GHQ categories

	Not stressed	Stressed
Samples	531	229
	Mean scores	
Good teaching	3.50	3.33
Clear goals	3.49	3.32
Generic skills	3.60	3.43
Appropriate assessment	3.22	3.20
Appropriate workload	3.21	2.93

Boxplots showing the distributions of the CEQ subscales by the GHQ categories are as in Figure 9.32 on page 258. There seem to be differences in the scores of CEQ subscales between the Stressed and Not Stressed groups. Those in the Stressed group appear to have slightly lower CEQ scores than those in the Not Stressed group.

To examine the relationship between course experience and general health, the mean scores of each of the CEQ subscales are compared between the two GHQ categories. Welch's two sample t-test is utilized, and the results are presented in Table 9.52 on page 257. Where there is a significant difference between the two means, the relationship is suggested to be significant.

Table 9.52: T-test on CEQ subscales by GHQ categories.

	P-value	df	95% CI	
GT	0.0013	437.77	0.064	0.263
CG	0.00015	413.98	0.083	0.257
GS	0.00078	413.65	0.072	0.273
AA	0.58	450.15	-0.067	0.120
AW	0.000017	392.51	0.152	0.401

Results indicate significant differences between the scores of four of the five subscales, ie. all except AA. These suggest that whether a participant falls in the category of Stressed or Not Stressed plays a part in how he or she experiences the training programme, as far as the four aspects of CEQ are concerned. The only one aspect where significance is not indicated is appropriate assessment, which relates to how the participants view their assessment. This seems to suggest that participants who are attending training while they are being stressed (for whatever reasons) experience the training programme less positively than their colleagues who are not stressed.

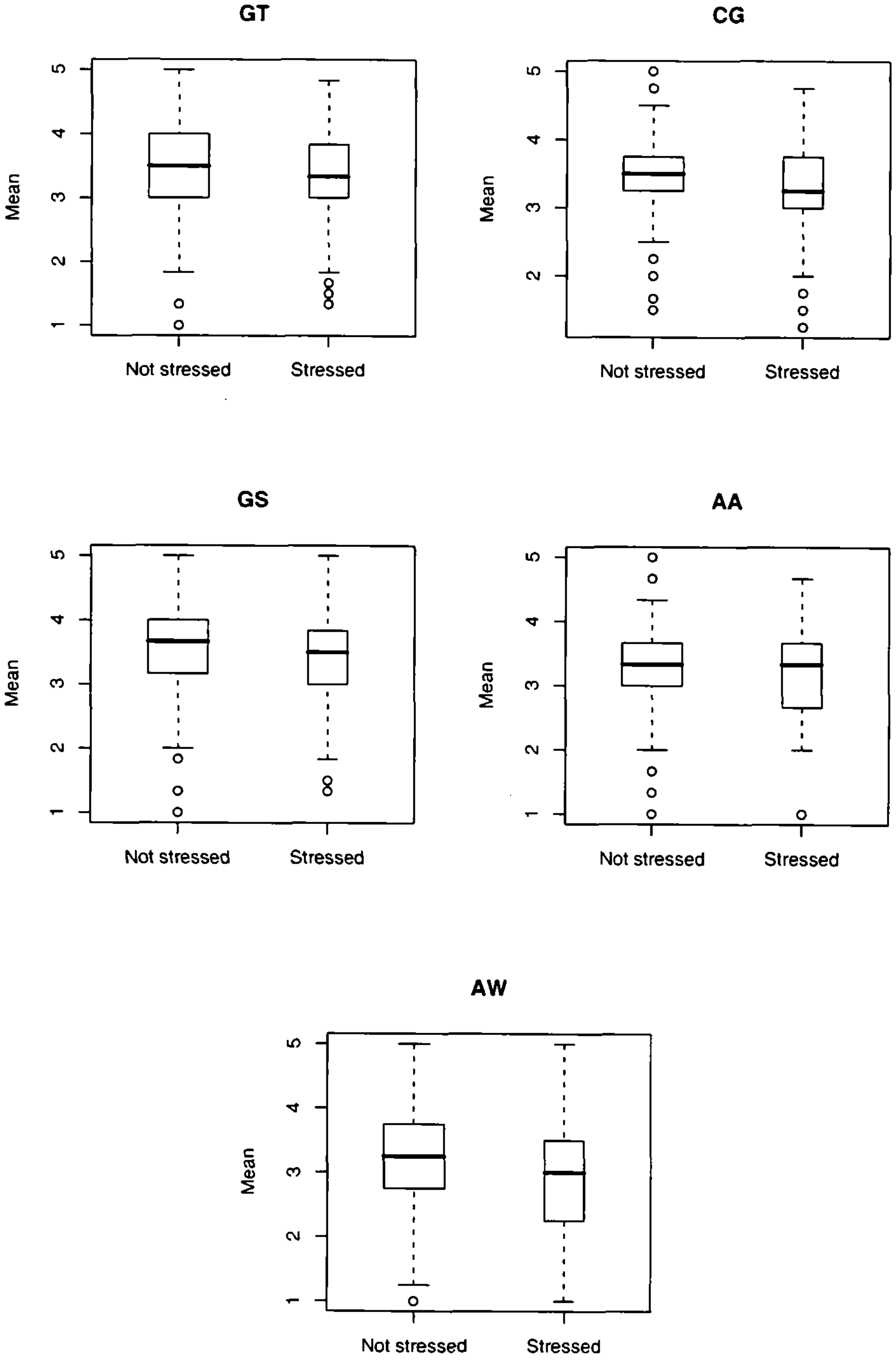


Figure 9.32: CEQ subscales by GHQ categories

Chapter 10

Model Comparison

The confirmatory factor analysis carried out in Chapter 6 tested whether the proposed model fit the data. For the Learning model, the finding was that the proposed models did not fit the data nicely. This is not a big surprise since that was an initial step in testing a newly developed model.

Testing of model fit can also be done by comparing the covariance matrix of the suggested model to the observed covariance matrix of the data. For single sample tests of dispersion, Krzanowski and Marriott [57] list several as follows:

- $H_0 : \Sigma = \Sigma_0$, μ unknown,
- $H_0 : \Sigma = k\Sigma_0$, k and μ unknown, and
- $H_0 : \Sigma$ is diagonal, μ is unknown,

where Σ is the observed covariance matrix and Σ_0 represents the hypothesized covariance structure.

In some situations there are competing hypothetical models available, and it is of interest to determine whether one, both, or neither are supported by the data. In this chapter, we adapt some ideas from Goldstein and Wooff ([36], chap. 9), to provide a graphical method for comparing a pair of hypothetical models.

In [36], methods of comparison are applied to alternative specifications for the prior variance matrix for a Bayesian analysis over a vector of random variables. In this chapter we will be concerned with comparing the variance matrices estimated given two differing SEM hypotheses. The two differing hypotheses are represented by Σ_1 and Σ_2. The eigenstructure of the compound matrix $K = \Sigma_2^{-1}\Sigma_1$ is then examined.

Before that, we will first have a look at the notation of the SEM in section 10.1, followed by a section on the estimation of the hypothesized covariance matrix. The method of the variance model comparison is explained in a section 10.3, followed by a section on the application of the method to the Learning data (Section 10.4). The analysis of the results of the method is discussed from section 10.5 onwards.

10.1 Notations

The notation for the structural equation model used in this thesis is mostly following Bollen [12]. The idea was first introduced in Section 3.8 on page 53. Bollen relies on the structural equation model notation developed by Joreskog [43,44], Wiley [95] and Keesling [50]. A full model of a structural equation contains random variables, structural parameters, and sometimes nonrandom variables. It has two major subsystems, namely (i) the latent variable model, and (ii) the measurement model.

10.1.1 The Latent Variable Model

A latent variable model is also known as the *structural model* or the *structural equation*. It represents the relationship between the latent variables. A latent variable corresponds to a concept, thus it is not observed. It is measured by one or more indicators which are measurable and observed. The relationship between the indicators and the latent variable they measure is the other subsystem of a structural equation model, namely the measurement model.

$$\eta = \beta\eta + \Gamma\xi + \zeta \tag{10.1}$$

Equation 10.1 represents the general matrix representation of the structural equation for the latent variable model. The first variable, η, is the vector size $m \times 1$ of the latent endogenous variables, where m is the number of the endogenous variables. The second variable, β, is the coefficient matrix size $m \times m$ for the latent endogenous variables. The third variable, Γ, is a $m \times n$ matrix of the coefficients for the latent exogenous variables, where n is the number of the exogenous variables. The fourth variable, ξ, is an $n \times 1$ vector of the exogenous latent variables. The last variable, ζ, is the vector size $m \times 1$ of the disturbances, or errors in the equations.

10.1.2 The Measurement Model

The measurement model provides the link between the latent variables and the observed variables. Since there are two types of latent variables (endogenous and exogenous) in the structural model, there are two different, but similar equations (Equations 10.2 and 10.3).

$$x = \Lambda_x \xi + \delta \tag{10.2}$$

$$y = \Lambda_y \eta + \epsilon \tag{10.3}$$

x and y are vectors of the observed indicators for ξ (exogenous) and η (endogenous) variables respectively. The Λ matrices are the coefficients relating x or y to their respective latent variables. The measurement errors for the observed variables are represented by the δ and ϵ vectors.

10.2 Estimation

In a single sample hypothesis testing procedure on the dispersion, a common hypothesis to be tested is the following:

$$\Sigma = \Sigma(\theta)$$

where Σ is the population covariance matrix of y (indicators of latent endogenous) and x (indicators of latent exogenous), and $\Sigma(\theta)$ is the implied covariance matrix, written as a function of model parameters in θ ([12], p. 325). We have

$$\begin{aligned}\Sigma(\theta) &= \begin{bmatrix}\Sigma_{yy}(\theta) & \Sigma_{yx}(\theta)\\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta)\end{bmatrix} \qquad (10.4)\\ &= \begin{bmatrix}\Lambda_y(I-\beta)^{-1}(\Gamma\Phi\Gamma'+\Psi)[(I-\beta)^{-1}]'\Lambda_y'+\Theta_\epsilon & \Lambda_y(I-\beta)^{-1}\Gamma\Phi\Lambda_x'\\ \Lambda_x\Phi\Gamma'[(I-\beta)^{-1}]'\Lambda_y' & \Lambda_x\Phi\Lambda_x'+\Theta_\delta\end{bmatrix}\end{aligned}$$

with the following entries:

Λ_y : coefficients relating y to η.

Λ_x : coefficients relating x to ξ.

β : coefficient matrix of latent endogenous (η).

Γ : coefficient matrix of latent exogenous (ξ).

Φ : covariance matrix of ξ.

Ψ : covariance matrix of ζ.

Θ_ϵ : covariance matrix of ϵ.

Θ_δ : covariance matrix of δ.

$\Sigma(\theta)$ is the hypothesized covariance matrix for the population. It is entirely based on the parameters of the hypothesized model and is independent of dataset. The values are unknown, and so must be estimated from the sample. Different samples produce different estimated matrices. In the R software, this estimated hypothesized covariance matrix is symbolised as C. In a structural equation modeling analysis, the model fit is tested by looking at the difference between the sample (observed) covariance matrix (S) and the matrix C.

10.3 Variance Model Comparison

When there are two competing hypothetical models, the hypotheses can be written as follows:

$$H_1 : \underline{X} \sim (\mu, \Sigma_1)$$

$$H_2 : \underline{X} \sim (\mu, \Sigma_2)$$

where $\underline{X}$ is a vector of measurements of k random variables, and Σ_1 and Σ_2 are $k \times k$ positive definite matrices. We would like to know which of the two covariance matrices resembles the sample covariance matrix, in which case it indicates the hypothesized structure fits the data. The possible results are:

- only one of the hypothesized structure fits the data,
- neither of the hypothesized structures fits the data, and
- both of the hypothesized structures fit the data.

The proposed method of model comparison involves solving the eigenstructure problem of the compound matrix as follows:

1. The compound matrix $K = \Sigma_2^{-1}\Sigma_1$ is formed and its eigenstructure is determined. There will be k eigenvectors and k eigenvalues ($\Lambda = diag[\lambda_1, ..., \lambda_k]$).

2. Let Σ_2 have normalized eigenvectors $Q = [\underline{q}_1, ..., \underline{q}_k]$ with eigenvalues $\Psi = diag[\psi_1, ..., \psi_k]$ (ie. $\Sigma_2 Q = Q\Psi$). It can be shown that $QQ^T = Q^T Q = I_k$ and $\Sigma_2 = Q\Psi Q^T$.

3. We solve the generalized eigenstructure problem of:

$$\Sigma_1 \underline{z} = \lambda \Sigma_2 \underline{z} \tag{10.5}$$

by writing

$$\Sigma_2^{-1}\Sigma_1 \underline{z} = \lambda \underline{z},$$

$$\text{ie.} \quad K\underline{z} = \lambda \underline{z}.$$

(a) Since $\Sigma_2 = Q\Psi Q^T$, (10.5) can be written as:

$$\Sigma_1 \underline{z} = \lambda \, Q\Psi Q^T \, \underline{z}. \tag{10.6}$$

(b) Then both sides of (10.6) are multiplied with $\Psi^{-\frac{1}{2}}Q^T$ giving:

$$\Psi^{-\frac{1}{2}}Q^T\Sigma_1\underline{z} = \lambda\Psi^{\frac{1}{2}}Q^T\underline{z}.$$

(c) Then a further multiplication to the left side of the equation;

$$\Psi^{-\frac{1}{2}}Q^T\Sigma_1\ Q\Psi^{-\frac{1}{2}}\Psi^{\frac{1}{2}}Q^T\ \underline{z} = \lambda\Psi^{\frac{1}{2}}Q^T\underline{z}, \qquad (10.7)$$

where $Q\Psi^{-\frac{1}{2}}\Psi^{\frac{1}{2}}Q^T = I_k$.

(d) Letting

$$\underline{y} = \Psi^{\frac{1}{2}}Q^T\underline{z}, \qquad (10.8)$$

we may write (10.7) as

$$\Psi^{-\frac{1}{2}}Q^T\Sigma_1 Q\Psi^{-\frac{1}{2}}\ \underline{y} = \lambda\underline{y}. \qquad (10.9)$$

(e) Letting $V = \Psi^{-\frac{1}{2}}Q^T\Sigma_1 Q\Psi^{-\frac{1}{2}}$, the problem of (10.5) has become solving $V\underline{y} = \lambda\underline{y}$. Note: the eigenvalues of matrix V are equal to the eigenvalues of matrix K, namely $\lambda_1, \lambda_2, ..., \lambda_k$.

(f) After solving $V\underline{y} = \lambda\underline{y}$, we find $Z = Q\Psi^{-\frac{1}{2}}Y$ from Equation (10.8), where Y is matrix of normalized eigenvectors of V, and Z is the matrix of eigenvectors in (10.5).

(g) From Equation 10.9:

$$\Psi^{-\frac{1}{2}}Q^T\Sigma_1 Q\Psi^{-\frac{1}{2}}\ Y = Y\Lambda. \qquad (10.10)$$

Pre-multiplying by Y^T:

$$\begin{aligned} Y^T\Psi^{-\frac{1}{2}}Q^T\Sigma_1 Q\Psi^{-\frac{1}{2}}\,Y &= Y^TY\Lambda. \\ \text{But}\quad Y^T\Psi^{-\frac{1}{2}}Q^T &= Z^T, \\ \text{and}\quad Q\Psi^{-\frac{1}{2}}Y &= Z. \\ \text{Therefore}\quad Z^T\Sigma_1 Z &= Y^TY\Lambda \qquad (10.11) \\ &= \Lambda. \end{aligned}$$

Meanwhile

$$\begin{aligned} Z^T\Sigma_2 Z &= Y^T\Psi^{-\frac{1}{2}}Q^T\Sigma_2 Q\Psi^{-\frac{1}{2}}Y. \\ \text{But}\quad Q^T\Sigma_2 Q &= \Psi. \\ \text{Therefore}\quad Z^T\Sigma_2 Z &= Y^T\Psi^{-\frac{1}{2}}\Psi\Psi^{-\frac{1}{2}}Y, \qquad (10.12) \\ &= Y^TY = I. \end{aligned}$$

We may summarize Equations 10.11 and 10.12 as follows:

$$Z^T\Sigma_1 Z = \Lambda, \quad \text{or} \quad \underline{z}_j^T\Sigma_1\underline{z}_j = \lambda_j \qquad (10.13)$$

$$Z^T\Sigma_2 Z = I_k, \quad \text{or} \quad \underline{z}_j^T\Sigma_2\underline{z}_j = 1 \qquad (10.14)$$

where the matrix $\Lambda = \text{diag}[\lambda_1, ...\lambda_k]$. We can also show that

- $\Sigma_1 = (Z^T)^{-1}\Lambda Z^{-1}$,
- $\Sigma_2 = (Z^T)^{-1}Z^{-1}$.

4. Define $W_j = \underline{z}_j^T(\underline{X} - \mu)$. This is a random variable with the following mean and variance.

$$E(W_j) = \underline{z}_j^T(E(\underline{X}) - \mu) = \underline{z}_j^T(\mu - \mu) = \underline{z}_j^T(0) = 0 \qquad (10.15)$$

$$\begin{aligned} Var(W_j) &= E[(W_j)^2] - [E(W_j)]^2 \\ &= E[(W_j)^2] - 0 \\ &= E[\underline{z}_j^T(\underline{X} - \mu)]^2 \\ &= E[\underline{z}_j^T(\underline{X} - \mu)(\underline{X} - \mu)^T\underline{z}_j] \\ &= \underline{z}_j^T E[(\underline{X} - \mu)(\underline{X} - \mu)^T]\underline{z}_j \\ &= \underline{z}_j^T Var(\underline{X})\underline{z}_j \end{aligned} \tag{10.16}$$

$Var(\underline{X})$ in (10.16) is equal to Σ_1 under hypothesis 1 (H_1) or Σ_2 under hypothesis 2 (H_2). Therefore, from (10.13) and (10.14), W_j is distributed as follows:

$$W_j \sim (0, \lambda_j) \qquad \text{under } H_1\text{, or} \tag{10.17}$$

$$W_j \sim (0, 1) \qquad \text{under } H_2. \tag{10.18}$$

If $\underline{X}$ is multivariate Normal, then its linear transformation W_j is similarly distributed, that is:

$$W_j \sim N(0, \lambda_j) \qquad \text{under } H_1\text{, or} \tag{10.19}$$

$$W_j \sim N(0, 1) \qquad \text{under } H_2. \tag{10.20}$$

$W_1, W_2, ..., W_k$ is a set of variables which are linearly transformed from a set of random variables $\underline{X}$ $(X_1, X_2, ..., X_k)$. While $\underline{X}$ is a set of correlated variables that has different structures between hypothesis 1 and hypothesis 2, $W_1, W_2, ..., W_k$ is a set of orthogonal components uncorrelated among each other (Cov(W_i, W_j) = 0, for all $i \neq j$). Furthermore, the structures of the two hypotheses are similar, and being uncorrelated, their variance implications may be assessed separately. These components differ only in variation, so differences in hypothesis may be explored fully through these variances.

10.3.1 Interpreting the Variance Comparison

The eigenvalues of the matrices $V = \Psi^{-\frac{1}{2}} Q^T \Sigma_1 Q \Psi^{-\frac{1}{2}}$ and $K = \Sigma_2^{-1}\Sigma_1$ are equal. Examining the individual eigenvalue λ_i of matrix V is the same as examining the eigenvalue of the original matrix K (Equation 10.5). Each eigenvalue shows how much larger or smaller is the variance of one component under a model compared to the variance of the same component under the other model. Equations 10.17 and 10.18 indicate that when the variance of W_j is fixed to 1 under hypothesis 2 (H_2), the following are true:

$\lambda_j > 1$ indicates W_j has larger variance under model H_1 than under model H_2;

$\lambda_j = 1$ indicates that $\text{Var}(W_j)$ is the same under both models;

$\lambda_j < 1$ indicates that W_j has lower variance under model H_1 than under model H_2.

10.3.2 Residual Analysis

We would like to examine the distributions of residuals under different hypotheses (models). Two covariance matrices will represent 2 different models. The residuals for examination are standardised as (10.21) and (10.22).

Let

$$R1_j = (z_j^T(\underline{X} - \mu))/\sqrt{\lambda_j} \tag{10.21}$$

and

$$R2_j = z_j^T(\underline{X} - \mu) \tag{10.22}$$

be the standardized residuals under the two hypotheses, where

$$E(R1_j) = 0, \quad Var(R1_j) = 1 \quad \text{under} \quad H_1 \tag{10.23}$$

$$E(R2_j) = 0, \quad Var(R2_j) = 1 \quad \text{under} \quad H_2 \tag{10.24}$$

and let the following be the observed values of $R1_j$ and $R2_j$:

$$r1_j = (\underline{z}_j^T(\underline{x} - \mu))/\sqrt{\lambda_j} \tag{10.25}$$

$$r2_j = \underline{z}_j^T(\underline{x} - \mu). \tag{10.26}$$

Equation 10.25 represents the observed value of standardised residuals under hypothesis 1 while (10.26) represents those under hypothesis 2. Both standardised residuals are expected to have zero mean and unit variance by conditions (10.23) and (10.24). If X is Normal, then these residuals should also be Normal.

10.3.3 Interpreting the Residuals

The distributions of both of the observed residuals $r1_j$ and $r2_j$ are assessed. Several approaches are used namely the following:

1. Graphical comparisons are possible using several different plots. First, we would like to present the discrepancies between the variances under the two hypotheses, in which case a plot of the eigenvalues (λ_j) is appropriate. Secondly, the distributions of the residuals may also be graphically presented to indicate their locations and dispersions.

2. It is also of interest to look at an overall measure of descrepancy between Σ_1 and Σ_2. Some possibilites are as follows:

   - λ_1 represents the maximum discrepancy for all $\lambda > 1$, assuming $\lambda_1 > 1$. On the other hand, λ_k represents the maximum discrepancy for all $\lambda < 1$, assuming $\lambda_k < 1$. In other words, λ_1 represents the highest variance difference for any liner combination of the original variables, relative to hypothesis 2 having variance 1.

   - $\tau = \frac{1}{k}\Sigma_{i=1}^{k}|log\lambda_i|$. If all λ_j were equal to 1, which indicates that all $Var(W_j)$ are the same under both hypotheses, τ would equal zero. On the other hand, if all λ_j were either much greater than or much less than 1, τ would be large, indicating variances which are highly different.

     In the same spirit, the value of e^{τ} is also calculated. A τ value of zero, which indicates equal variances under both hypotheses, would result in a e^{τ} of 1, the same value at which an eigenvalue indicates equal variances.

   - Further research could also look at the probability distribution for eigenvalues $\lambda_1, ...\lambda_k$, and could also bootstrap the summary statistics. Keramidas, Devlin and Gnanadesikan [51] propose a graphical procedure for comparing the principal components (PCs) of several covariance matrices, and propose a hypothesis test of equality of PCs. These methods are not, however, based on a joint decomposition of the variance matrices, but on "averaging" the PCs. Flury [29] proposes a basic test of equality of PCs.

3. It may be reasonable to assume that the residuals should be Normally distributed approximately. If so, then each $r1_j$ and $r2_j$ can be tested for a standard Normal distribution. This may be accomplished by running a Kolmogorov-Smirnov (K-S) one-sample test. Standardised residuals have expectation of zero mean and unit variance. The usual K-S test is simply one of Normality, but the test applied here is a more stringent Normality test where the hypothesis tested is not general Normality but one with mean zero and variance one.

10.4 Application of the Method to Learning Data

In applying this method to the Learning data, we need to estimate the variance matrices for the different hypotheses, and so the comparison relates to these estimates, and less directly to the underlying hypotheses. In other words this method is exploratory. It is applied to the Learning data based on the new Learning model. The original Learning model with 30 indicators has been modified, leaving only 18 indicators, as explained in Section 6.3.2 on page 109. Σ_1 is taken to be the hypothesized covariance matrix of the 18 indicators. Σ_2 is taken to be the model-implied covariance matrix, obtained from the SEM analysis on the data.

The matrix Σ_1 reflects the hypothesis that the variables, or indicators of the latent factor, are independent of each other. The matrix is diagonal, and the diagonal entries are estimated from the sample. As there are 18 variables in the model, Σ_1 is 18×18 as in (10.27).

$$\Sigma_1 = \begin{bmatrix} \sigma^2_{1.1} & 0 & .. & 0 \\ 0 & \sigma^2_{2.2} & .. & 0 \\ .. & .. & .. & .. \\ 0 & 0 & .. & \sigma^2_{18.18} \end{bmatrix} \tag{10.27}$$

The model-implied covariance matrix Σ_2 (which is also 18×18) is an estimation of $\Sigma(\theta)$, the population model-implied covariance matrix (10.4). In conventional SEM analysis using the R package, this covariance matrix is produced through the maximum likelihood minimizing function [32] and then compared with the sample (observed) covariance matrix S to test model fit. The values of this matrix depend on the data used in creating it. Since the pretest data is going to be used in the application of this method, the same data is chosen to create Σ_2, instead of the posttest data.

Using the method described in Section 10.3, eighteen original random variables have now been transformed into a set of eighteen new random variables, in the form of the components $W_j, j = 1$ to 18. These new random variables are nicer than the original random variables in the sense that they are orthogonal, and have expectation 0 and variance either λ_j or 1, depending on the hypothesis.

10.4.1 Implied Covariance Matrix

The matrix $\Sigma(\theta)$ as presented in (10.4) represents the implied covariance matrix for the general structural equation model. The Learning model used for this method is presented as Figure 6.6 on page 110. Model parameters include eighteen indicators for endogenous variables, three endogenous variables and one exogenous variable. Since the endogenous variables do not affect each other, β is zero (($\beta_{ij} = 0$, for all i and j) [12], p. 15) thus reducing the structural equation for the latent measurement model of (10.1) on page 260 to (10.28).

$$\eta = \Gamma\xi + \zeta \tag{10.28}$$

The relevant matrices of the structural model (10.28) are the following:

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}, \quad \Gamma = \begin{bmatrix} 1 \\ \gamma_{21} \\ \gamma_{31} \end{bmatrix}, \quad \xi = \begin{bmatrix} \xi_1 \end{bmatrix}, \quad \zeta = \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{bmatrix},$$

$$\phi = \begin{bmatrix} Var(\xi_1) \end{bmatrix}, \quad \psi = \begin{bmatrix} Var(\zeta_1) & 0 & 0 \\ 0 & Var(\zeta_2) & 0 \\ 0 & 0 & Var(\zeta_3) \end{bmatrix}. \tag{10.29}$$

The η matrix shows the three latent endogenous variables. The single exogenous variable in the model is presented in the ξ matrix, while its direct effects on the endogenous variables are presented as coefficients in the Γ matrix. In the Γ matrix the first coefficient is fixed to 1 to identify the exogenous variable. The ζ matrix shows the latent errors associated with the three endogenous variables. The ϕ matrix is the covariance matrix of the exogenous variable. Since there is only one exogenous variable, this matrix is a scalar that equals the variance of ξ_1. The ψ matrix is the covariance of the latent errors. The matrix is diagonal because the errors are assumed to be uncorrelated. The measurement part of the Learning model is presented as (10.30):

$$y = \Lambda_y \eta + \epsilon \tag{10.30}$$

Matrices relevant to the measurement model are the following:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_{18} \end{bmatrix}, \quad \Lambda_y = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{2.1} & 0 & 0 \\ \vdots & \vdots & \vdots \\ \lambda_{6.1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{8.2} & 0 \\ \vdots & \vdots & \vdots \\ 0 & \lambda_{12.2} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{14.3} \\ \vdots & \vdots & \vdots \\ 0 & 0 & \lambda_{18.3} \end{bmatrix}, \quad \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix},$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_{18} \end{bmatrix}, \quad \theta_\epsilon = \begin{bmatrix} Var(\epsilon_1) & 0 & \cdots & 0 \\ 0 & Var(\epsilon_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & Var(\epsilon_{18}) \end{bmatrix}. \tag{10.31}$$

The y matrix contains the eighteen observed indicators of the endogenous variables (η). Their relations to the three endogenous are presented as coefficients in the Λ_y matrix. The ϵ matrix shows the errors associated with the measurement of the indicators, and the θ_ϵ matrix is the covariance matrix of the errors. The errors are assumed to be uncorrelated, thus the matrix is diagonal with error variances as diagonal entries.

In the Learning model, only endogenous variables are indicated by observed variables. The exogenous variable (namely Learning) is explained indirectly by the same observed variables through the three endogenous variables. In other words, this model has"y" variables but no "x" variables. Consequently, $\Sigma_{yx}(\theta)$, $\Sigma_{xy}(\theta)$ and $\Sigma_{xx}(\theta)$ in (10.4) are all non-existent. Only the upper-left quadrant is left, as (10.32) below.

$$\Sigma(\theta) = \Big[\Lambda_y(I-\beta)^{-1}(\Gamma\Phi\Gamma' + \Psi)[(I-\beta)^{-1}]'\Lambda_y' + \Theta_\epsilon\Big] \tag{10.32}$$

Replacing the matrices (10.29) and (10.31) of the model parameters (and zero β matrix) into (10.32) produces an 18×18 matrix of $\Sigma(\theta)$. An estimate of matrix (10.32) from the data is used as Σ_2 for the application of this method in this chapter.

10.5 Analysis

Between Σ_1 and Σ_2, it is expected beforehand that the data will fit Σ_2 better. This is due to the fact that Σ_2 is the model implied covariance matrix whose values are estimated from the data. In the case of Σ_1, the diagonal values are also estimated from the sample, but the independence assumption among the indicators is hypothetical and not derived from the data.

Analysis is carried out separately for each dataset (pretest and posttest). The method is used on the pretest data and repeated for the posttest data. There are two main steps involved. In the first step, SEM analysis (Section 3.8 on page 53) is carried out on the data. This analysis produces a matrix which estimates $\Sigma(\theta)$, the implied covariance matrix (Equation 10.4). This matrix is renamed as Σ_2.

In the second step, the original eighteen indicators of the Learning model are linearly transformed into a set of eighteen orthogonal components. The input to this step is the Σ_1 and Σ_2 matrices. Following these two steps, this method produces the following outputs from each dataset:

1. a matrix of eigenvalues $\Lambda = diag[\lambda_1, \lambda_2, ..., \lambda_{18}]$, corresponding to the orthogonal components W_j. The eigenvalues are produced after solving the compound matrix of (10.5).

2. a matrix of eigenvectors $Z = [\underline{z}_1, \underline{z}_2, ..., \underline{z}_{18}]$, related to the same equation (10.5). These eigenvectors are the coefficients for the linear combinations of components W_j over the original variables.

3. a matrix of residuals under hypothesis 1 (Equation 10.25), and

4. a matrix of residuals under hypothesis 2 (Equation 10.26).

10.6 Analysis on the pretest data

In this section, the pretest data is used to estimate Σ_1 and Σ_2. The former matrix is diagonal, with the sample covariances as the diagonal entries. The latter matrix is produced when the SEM analysis is carried out on the pretest data.

10.6.1 Eigenvalues

Following the application of this method on the pretest data, the eigenvalues of the components W_j are produced and presented in Table 10.1. A plot of the eigenvalues is presented in Figure 10.1.

The first fifteen eigenvalues are greater than 1, and the last three are less than 1. This shows that the first fifteen components have higher variance under hypothesis 1 ($H_1 : \underline{X} \sim (\mu, \Sigma_1)$) than they do under hypothesis 2 ($H_2 : \underline{X} \sim (\mu, \Sigma_2)$). The last

Table 10.1: Eigenvalues from the pretest data.

W_j	Eigenvalue	W_j	Eigenvalue	W_j	Eigenvalue
1	3.35	7	2.17	13	1.63
2	3.14	8	2.01	14	1.56
3	2.63	9	1.92	15	1.27
4	2.56	10	1.73	16	0.91
5	2.24	11	1.68	17	0.38
6	2.19	12	1.66	18	0.15

three W_js on the other hand have lower variance under the first model than they do under the second one. The sum of the absolute values of the log is 13.66, giving the value of τ equals 0.76. This value is far from zero, reflecting the variance differences between the two hypotheses. The value of e^{τ} is 2.14.

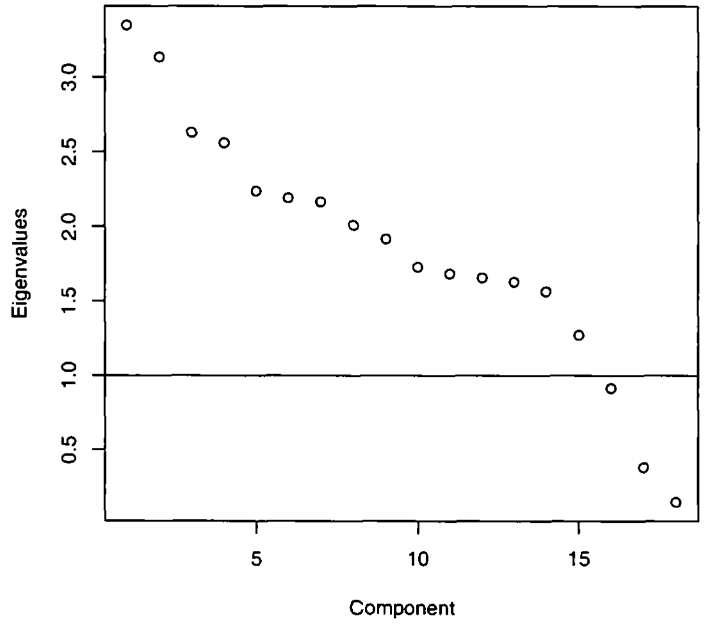


Figure 10.1: Eigenvalues from the pretest data

10.6.2 Eigenvectors

The corresponding eigenvectors are presented in Table 10.2 on page 276. The first component relates to the largest discrepancy between Σ_1 and Σ_2 for $\lambda > 1$, while the last component relates to the largest discrepancy for $\lambda < 1$. Relationships between these components and the variables could be summarized (rounded to 2 decimal places) as (10.33) and (10.34).

$$W_1 \approx -1.13Y_{13} + 1.04Y_{16} \tag{10.33}$$

$$W_{18} \approx -0.10Y_{13} - 0.12Y_{14} - 0.11Y_{15} \tag{10.34}$$

Equation 10.33 indicates that the first component W_1 is made up basically of variables Y_{13} and Y_{16}. The linear combination of these variables has the largest variance under hypothesis 1, assuming that the variance under hypothesis 2 is fixed to 1 for the pretest data. Since this component is associated with the largest eigenvalue, it also implies that this linear combination constitutes the largest discrepancy between hypothesis 1 and hypothesis 2 in this dataset.

Eigenvectors related to the last component, W_{18} (as presented in Table 10.1), indicate that this component is approximately an average of all the original variables, with larger components for Y_{13}, Y_{14}, and Y_{15}, as indicated by Equation 10.34. Among the components with lower variance under hypothesis 1 than under hypothesis 2 ($\lambda_j < 1$), this one has the largest discrepancy between the two hypotheses.

As listed in Table 6.17 on page 111, each of the variables Y_i refers to a subject area. The first component, W_1, is thus a linear combination of (i) the *importance* of the subject of **Economic Management** (Y_{13}), and (ii) the *importance* of the subject of **Social and Infrastructure Planning** (Y_{16}).

10.6.3 Standardized Observed Residuals of the Pretest Data

We begin by exploring the possibility that the residuals are Normally distributed, and with mean zero and variance one. Residuals from the pretest data are examined using histograms and normality plots in Figure 10.2. Both histograms centre on zero, but the shapes are different. R2 has a fatter distribution compared to R1, but R1

	Components																	
Y	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}	W_{11}	W_{12}	W_{13}	W_{14}	W_{15}	W_{16}	W_{17}	W_{18}
1	0.00	0.00	0.00	0.00	0.29	-0.01	0.01	-0.01	-0.00	-0.01	0.55	-0.70	-0.21	-0.00	0.14	-0.05	0.20	-0.04
2	0.00	0.00	0.00	0.00	0.09	-0.00	0.00	-0.00	-0.00	-0.00	0.01	0.01	0.01	0.00	-0.94	-0.07	0.16	-0.03
3	0.00	0.00	0.00	0.00	0.29	-0.01	0.01	-0.01	-0.00	-0.01	0.15	0.25	1.00	-0.01	0.18	-0.06	0.22	-0.04
4	0.00	0.00	0.03	0.00	-1.05	0.02	-0.02	0.01	0.00	0.00	-0.02	-0.02	-0.02	-0.00	0.08	-0.04	0.21	-0.05
5	0.00	0.00	0.00	0.00	0.29	-0.01	0.01	-0.01	-0.00	-0.02	-0.79	-0.20	-0.12	-0.00	0.12	-0.05	0.18	-0.04
6	0.00	0.00	0.00	0.00	0.28	-0.01	0.01	-0.01	-0.00	-0.01	0.23	0.75	-0.53	-0.01	0.15	-0.06	0.20	-0.04
7	-0.00	-0.00	0.38	-0.12	0.04	0.83	0.01	0.38	0.00	0.22	-0.00	-0.00	0.00	0.08	-0.01	0.22	-0.02	-0.08
8	-0.00	-0.00	0.22	-0.06	0.00	-0.10	-0.00	-1.00	0.08	0.54	-0.00	-0.00	0.00	0.15	-0.01	0.27	-0.02	-0.09
9	-0.00	-0.00	0.35	-0.11	0.01	-0.84	-0.11	0.60	0.00	0.28	-0.00	-0.00	0.00	0.10	-0.01	0.24	-0.02	-0.09
10	-0.00	-0.02	-0.97	0.20	-0.00	0.05	0.00	0.12	0.00	0.13	-0.00	-0.00	0.00	0.05	-0.00	0.18	-0.01	-0.08
11	-0.00	-0.00	0.08	-0.02	0.00	-0.02	-0.00	-0.11	-0.00	-0.63	0.01	0.00	0.00	0.38	-0.01	0.20	-0.01	-0.06
12	-0.00	-0.00	0.08	-0.02	0.00	-0.02	-0.00	-0.10	-0.00	-0.33	0.00	0.00	-0.00	-0.88	-0.02	0.27	-0.01	-0.07
13	-1.13	0.76	0.12	0.44	-0.00	-0.01	0.17	-0.01	-0.10	-0.01	-0.00	-0.00	-0.00	-0.01	0.00	-0.21	-0.08	-0.10
14	-0.01	-0.13	-0.21	-1.10	-0.02	-0.07	0.80	-0.02	-0.24	-0.02	-0.00	-0.00	-0.00	-0.01	0.01	-0.25	-0.08	-0.12
15	-0.04	-1.27	0.20	0.68	-0.00	-0.02	0.22	-0.01	-0.12	-0.01	-0.00	-0.00	-0.00	-0.01	0.00	-0.22	-0.08	-0.11
16	1.04	0.61	0.10	0.38	-0.00	-0.01	0.15	-0.00	-0.09	-0.01	-0.00	-0.00	-0.00	-0.01	0.00	-0.18	-0.07	-0.09
17	-0.00	-0.06	-0.06	-0.29	0.00	0.04	-0.80	-0.06	-0.43	-0.02	-0.00	-0.00	-0.00	-0.01	0.00	-0.18	-0.06	-0.07
18	-0.00	-0.05	-0.05	-0.21	0.00	0.02	-0.29	0.02	1.06	-0.06	-0.00	-0.00	-0.00	-0.02	0.01	-0.23	-0.07	-0.09

Table 10.2: Eigenvectors of the pretest data, shown as coefficients of the Y_is.

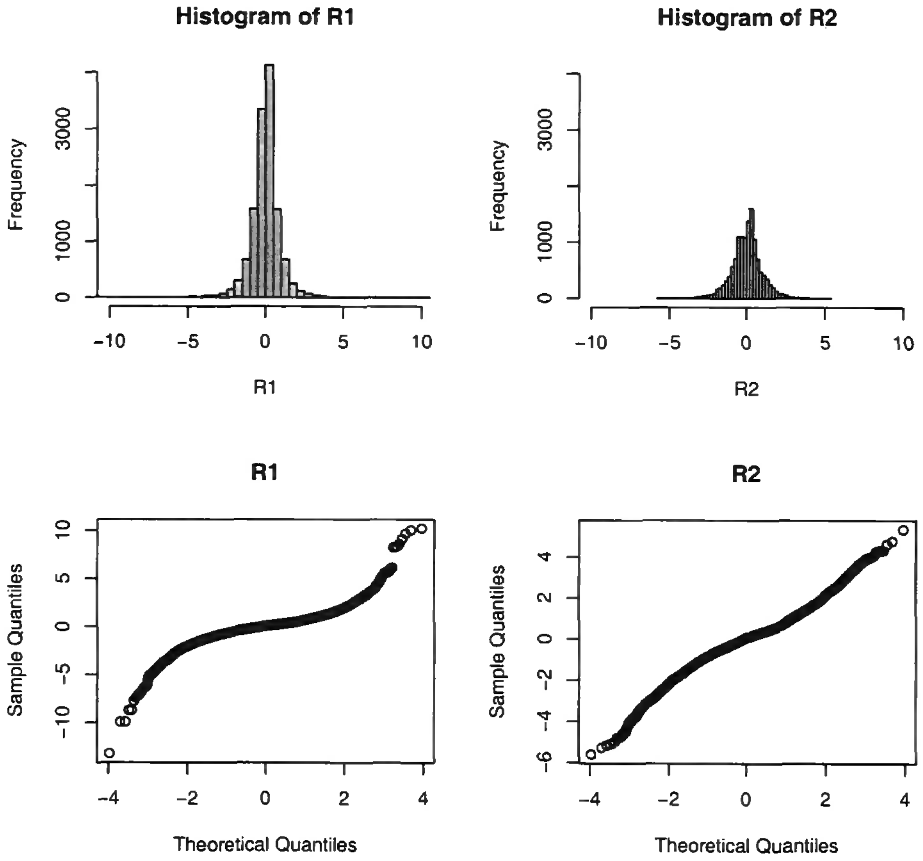


Figure 10.2: Histograms and normality plots of standardized residuals under H_1 (R1) and H_2 (R2) of the pretest data.

has longer tails on both sides. In terms of Normality, R2, the standardized residuals under H_2, seems to be approximating Normal distribution better than does R1. This condition is also shown by the QQ plots.

On conditions (10.23) and (10.24), each of the $R1_j$ and $R2_j$ is expected to have mean zero and variance one. To test them for these expectations, the Kolmogorov-Smirnov (K-S) test is utilized. The K-S test is a distribution-free test, used when we want to know whether observations are consistent with their being a sample from some specified continuous distribution [84]. In this particular work, this test is used to test the hypothesis that not only the residuals are Normally distributed, but that they have mean zero and variance one as well.

Table 10.3 shows the p-values of K-S tests on the hypothesis that the standard-

Table 10.3: P-values of the K-S test on $R1_j$ and $R2_j$ of pretest data.

j	$R1_j$	$R2_j$
1	0.000e+00	0.000e+00
2	0.000e+00	0.000e+00
3	2.176e-09	6.898e-07
4	1.081e-13	6.205e-11
5	7.946e-05	7.211e-02
6	0.000e+00	0.000e+00
7	2.555e-11	1.895e-13
8	3.523e-11	5.524e-07
9	3.888e-15	1.223e-11
10	6.375e-05	1.377e-04
11	1.101e-05	5.362e-03
12	3.638e-08	1.005e-03
13	5.631e-05	2.021e-03
14	2.366e-08	3.257e-04
15	3.280e-07	8.138e-03
16	1.617e-07	8.361e-07
17	1.284e-04	3.485e-02
18	4.795e-04	2.859e-02

ized residuals are distributed Normally with mean zero and standard deviation one (*N(0,1)*). Column 2 relates to residuals under hypothesis 1, while column 3 relates to residuals under hypothesis 2. Generally, all p-values imply evidence against the underlying residuals having a N(0,1) distribution.

Figure 10.3 shows an informal plot of the negative of the log p-values. Smaller p-values are represented by (larger) higher points on the plot, indicating components with more evidence against the null hypothesis of N(0,1). In other words, a residual which fits the hypothesis of N(0,1) better is indicated by a lower point on the plot. Components whose R1 residuals fit the N(0,1) better than their R2 is component 7 only. The conclusion is that neither model appears to generate N(0,1) residuals, but that model one is more abnormal in this regard.

It would be more natural to assess only whether the mean is zero and the variance is unity without reference to an underlying continous distribution. However, there do not appear to be satisfactory tests available in the literature, and generating them is outside the scope of this thesis. Figure 10.4 shows the distributions of the absolute residuals under the two hypotheses. We can see that generally, the distributions of

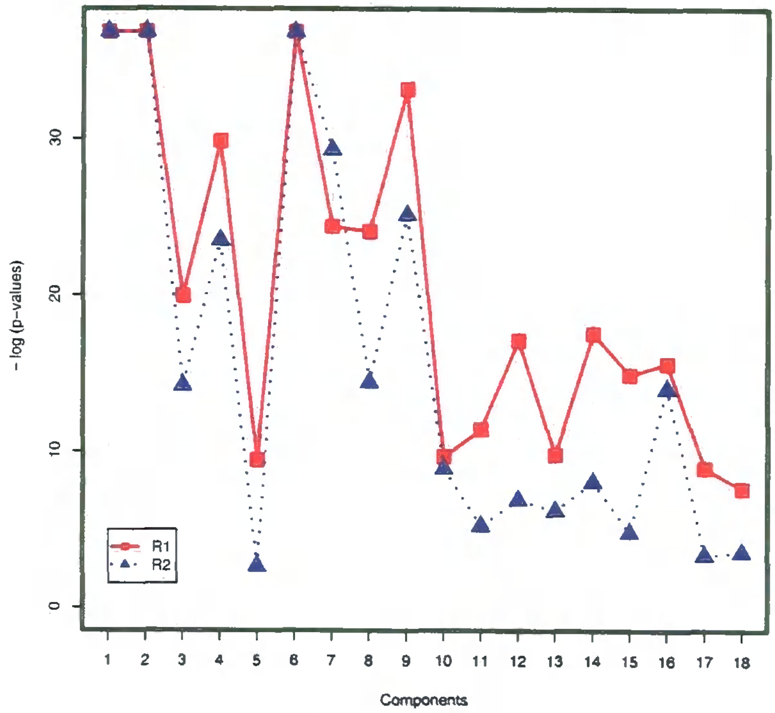


Figure 10.3: Plot of the negative log(p-values) of K-S test on pretest data. Small values imply greater consistency with a N(0,1) hypothesis.

the residuals under H_1 are wider than those under H_2. This indicates that the majority of the components have larger residuals under the first hypothesis than they do under the second one.

10.7 Analysis on the Posttest Data

In this section, the posttest data is used to estimate Σ_1 and Σ_2. Following the application of the variance model comparison method on the posttest data, we discuss the following results. The eigenvalues of the components W_j are presented in Table 10.4.

A total of sixteen eigenvalues from this data are greater than 1. The other two are much less than 1 as shown in Figure 10.5. This indicates that the first sixteen components have higher variance under hypothesis 1 ($H_1 : \underline{X} \sim (\mu, \Sigma_1)$) than they

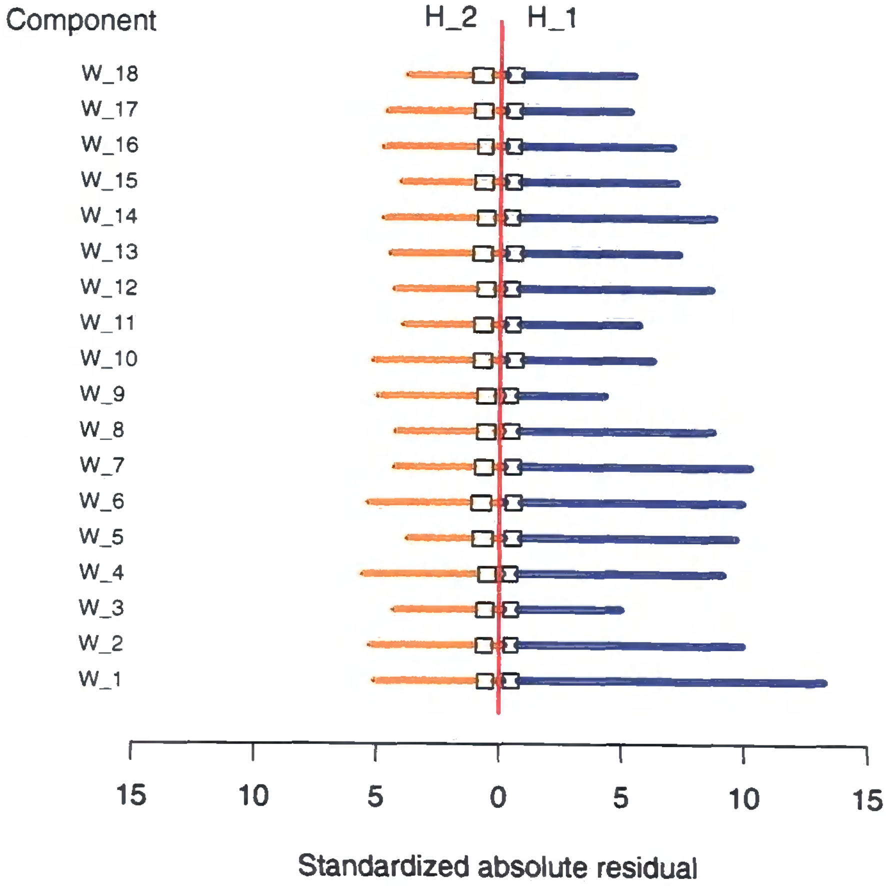


Figure 10.4: Distributions of the absolute pretest residuals by components.

Table 10.4: Eigenvalues from the posttest data.

W_j	Eigenvalue	W_j	Eigenvalue	W_j	Eigenvalue
1	3.33	7	2.17	13	1.70
2	3.00	8	2.10	14	1.57
3	2.76	9	2.04	15	1.47
4	2.31	10	1.95	16	1.25
5	2.29	11	1.73	17	0.49
6	2.20	12	1.72	18	0.13

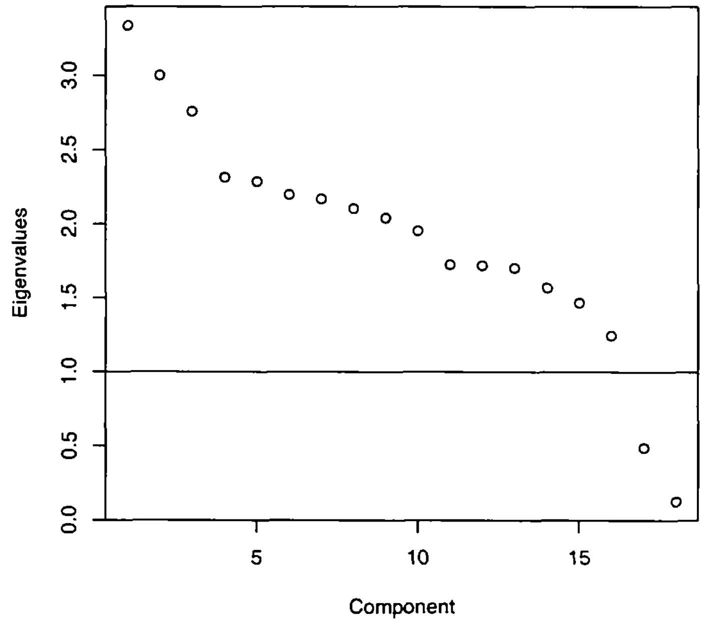


Figure 10.5: Eigenvalues from the posttest data

do under hypothesis 2 ($H_2 : \underline{X} \sim (\mu, \Sigma_2)$). The last two components have lower variance under the second model than they do under the first one. The sum of the absolute values of the log of the eigenvalues is 14.11, giving the value of τ equals to about 0.78. This value is just slightly more than the value for the pretest data, which is 0.76. The value of e^τ is 2.19, compared to 2.14 of the pretest data. These results of the posttest data are quantitatively similar to the results of the pretest data.

10.7.1 Eigenvectors

The eigenvectors from the posttest data are as in Table 10.5 on page 283. The first and the last components can be summarized as equations (10.35) and (10.36).

$$W_1 \approx 0.99Y_{18} \tag{10.35}$$

$$W_{18} \approx 1.00Y_1 \tag{10.36}$$

Equations (10.35) and (10.36) indicate that both of the components are made up of almost entirely a single variable each. The first component which has the largest eigenvalue is almost entirely made up of variable Y_{18}, while the last component is made up of variable Y_1. These are different to (10.33) and (10.34) of the pretest data, which is hard to explain.

The variable Y_{18} refers to to the *importance* of the **specific knowledge** targetted by the training programme (Please refer Table 6.17 on page 111). Equation 10.35 indicates that the component W_1 is a linear combination of this subject area almost in totality. The variable Y_1 meanwhile refers to the *knowledge* in the subject area of **Economic Management**. Equation 10.36 indicates that component W_{18} is a linear combination of this subject area, also almost entirely.

10.7.2 Standardized Observed Residuals of the Posttest Data

Residuals from the posttest data are standardized as in (10.21) and (10.22). The distributions are examined using histograms and normality plots (Figure 10.6). Just like the residuals of the pretest data, the posttest data residuals also centre on zero, but the distributions of R1 and R2 are different. The distribution of R1 is thin with long tails, while the distribution of R2 is fatter. Looking at the QQ plots we can see clearly that the distribution of R2 seems to be approximating the Normal distribution better than does R1. However, neither distribution appears Normal.

We have shown by (10.23) and (10.24) that each of the $R1_j$ and $R2_j$ is expected to be Normally distributed with mean zero and variance one. The Kolmogorov-Smirnov tests on this hypothesis produce p-values as in Table 10.6.

Figure 10.7 shows the negative log of the p-values. Generally it can be seen that there are more components whose R2 residuals fit the N(0,1) hypothesis better than their R1 residuals. This finding is similar to that of the pretest data. This is again emphasized by Figure 10.8, where we can see generally the residuals under H_2 are smaller than the residuals under H_1.

	Components																	
Y	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}	W_{11}	W_{12}	W_{13}	W_{14}	W_{15}	W_{16}	W_{17}	W_{18}
1	0.00	0.01	-0.02	0.00	0.02	-0.01	0.02	0.02	-0.03	0.00	0.04	-0.01	-0.01	-0.01	-0.03	-0.03	-0.02	1.00
2	0.00	-0.01	0.03	-0.01	-0.01	-0.05	-0.01	-0.01	0.05	0.00	-0.04	-0.04	-0.06	-0.04	-0.18	0.02	0.98	0.02
3	0.01	0.02	-0.05	0.03	-0.03	-0.01	-0.01	-0.01	-0.15	0.05	-0.63	0.02	0.00	0.01	-0.08	0.75	-0.05	0.04
4	0.00	0.01	-0.01	0.01	-0.03	0.00	-0.07	-0.04	-0.06	-0.07	0.27	0.05	0.12	-0.26	-0.89	0.12	-0.16	-0.03
5	0.01	0.02	-0.04	0.08	-0.14	0.03	-0.31	-0.14	-0.26	-0.23	0.62	-0.07	-0.10	0.09	0.27	0.51	0.07	0.00
6	0.00	-0.03	0.15	0.25	0.25	-0.05	0.47	0.00	-0.78	0.03	0.04	0.00	0.02	-0.01	0.00	-0.10	0.05	-0.04
7	0.00	0.00	-0.01	-0.02	0.03	-0.09	0.09	-0.01	0.04	0.01	0.01	-0.08	-0.74	0.59	-0.27	-0.02	-0.08	-0.02
8	0.01	0.04	-0.12	-0.17	0.13	-0.06	0.73	-0.29	0.38	0.00	0.24	-0.01	0.02	-0.10	0.07	0.30	0.00	0.00
9	0.00	0.00	0.01	0.01	0.01	-0.02	-0.05	0.05	-0.02	0.02	-0.03	0.23	-0.65	-0.71	0.13	-0.02	-0.04	-0.01
10	0.01	0.04	-0.08	-0.11	0.31	-0.04	-0.05	0.58	0.04	0.65	0.26	-0.08	0.02	0.01	0.02	0.20	0.01	-0.02
11	0.00	0.01	-0.01	-0.01	-0.01	0.15	0.02	-0.01	0.00	0.08	0.08	0.95	0.07	0.24	0.00	0.03	0.06	0.01
12	-0.02	-0.08	0.19	0.28	-0.41	0.02	-0.03	-0.48	0.01	0.68	0.04	-0.06	-0.02	-0.02	-0.01	-0.03	-0.02	0.01
13	0.06	0.23	-0.51	-0.63	-0.29	-0.01	0.01	-0.14	-0.37	0.16	-0.04	-0.02	-0.01	-0.01	-0.01	-0.14	0.03	-0.02
14	0.00	0.00	0.00	-0.18	0.73	0.00	-0.35	-0.54	-0.03	0.12	-0.05	0.01	0.00	0.00	0.00	-0.03	0.00	0.00
15	0.05	0.22	-0.39	0.34	0.10	0.80	0.05	0.00	0.05	0.01	-0.02	-0.12	-0.07	-0.01	-0.03	-0.04	0.04	-0.01
16	-0.07	-0.30	0.54	-0.52	-0.06	0.56	0.07	0.04	-0.07	0.00	0.02	-0.08	-0.05	0.00	-0.02	0.07	0.00	0.02
17	0.01	-0.89	-0.44	0.05	0.01	0.00	0.00	0.00	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00
18	0.99	-0.04	0.10	-0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 10.5: Eigenvectors of posttest data, shown as coefficients fo the Y_is.

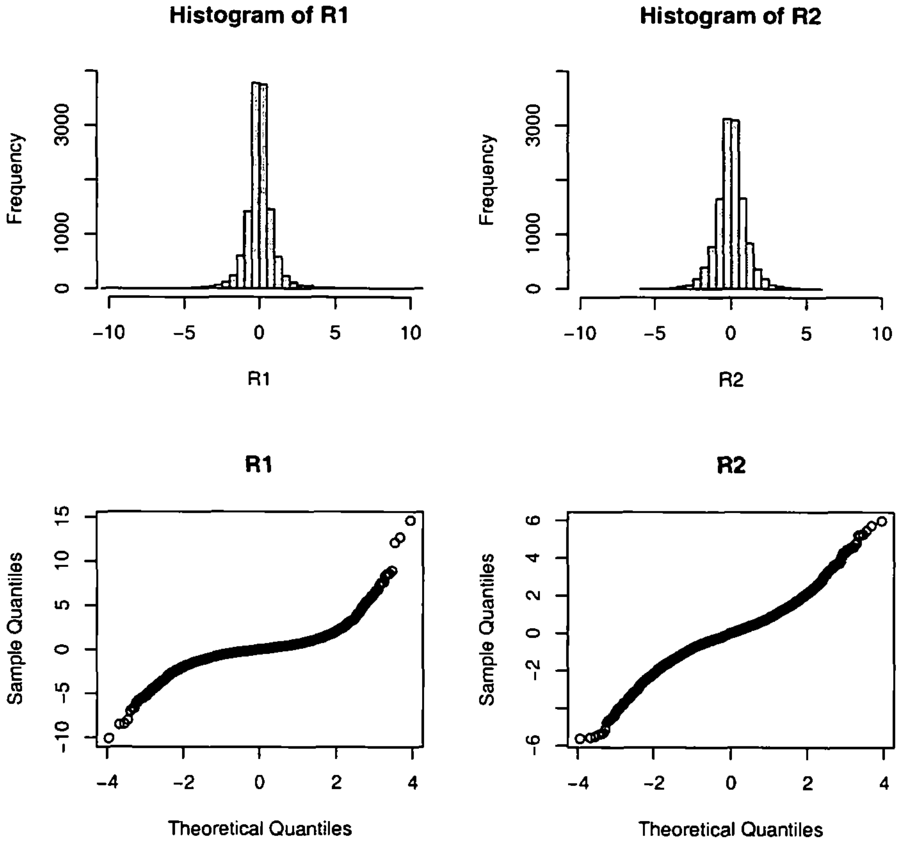


Figure 10.6: Histograms and normality plots of standardized residuals under H_1 (R1) and H_2 (R2) of the posttest data.

Table 10.6: P-values of the K-S test on $R1_j$ and $R2_j$ of posttest data.

j	$R1_j$	$R2_j$
1	0.000e+00	0.000e+00
2	0.000e+00	0.000e+00
3	0.000e+00	0.000e+00
4	1.593e-12	1.583e-11
5	7.867e-10	1.984e-07
6	1.586e-07	5.714e-06
7	2.227e-10	2.499e-09
8	0.000e+00	0.000e+00
9	1.110e-16	1.110e-15
10	3.259e-09	4.479e-08
11	1.533e-09	2.362e-08
12	1.154e-08	2.125e-07
13	2.248e-09	3.818e-08
14	1.106e-08	1.147e-04
15	1.622e-09	8.024e-06
16	8.662e-09	5.018e-07
17	8.062e-07	1.934e-02
18	2.303e-03	3.603e-02

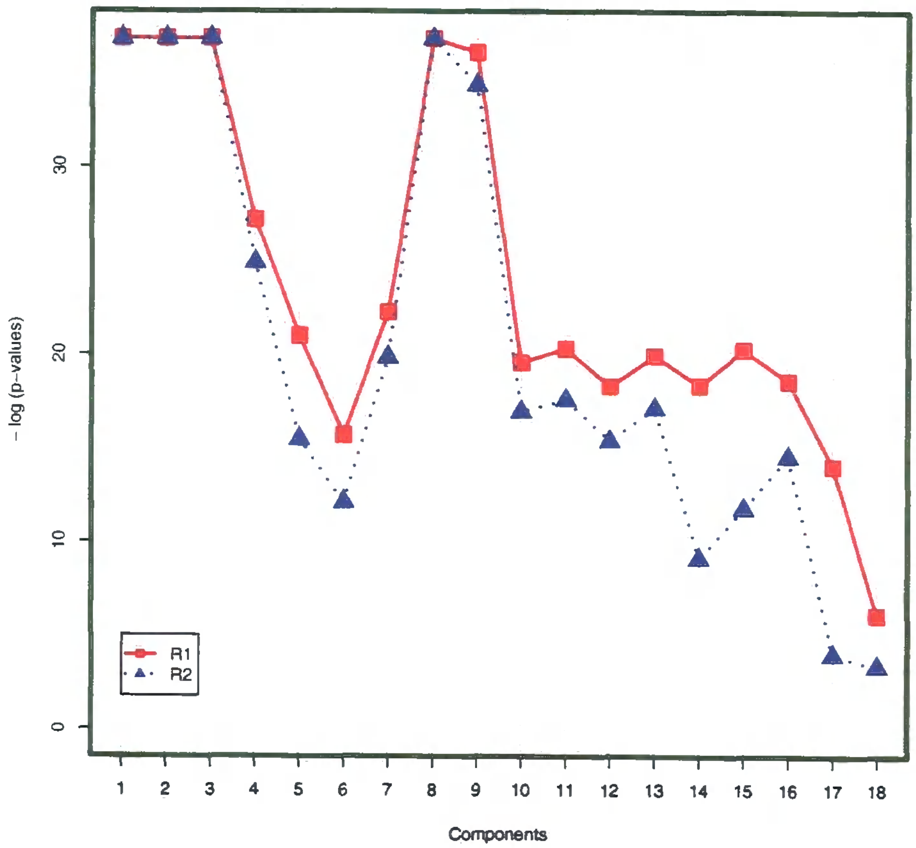


Figure 10.7: Plot of the negative log(p-values) of K-S test on posttest data. Small values imply greater consistency with a N(0,1) hypothesis.

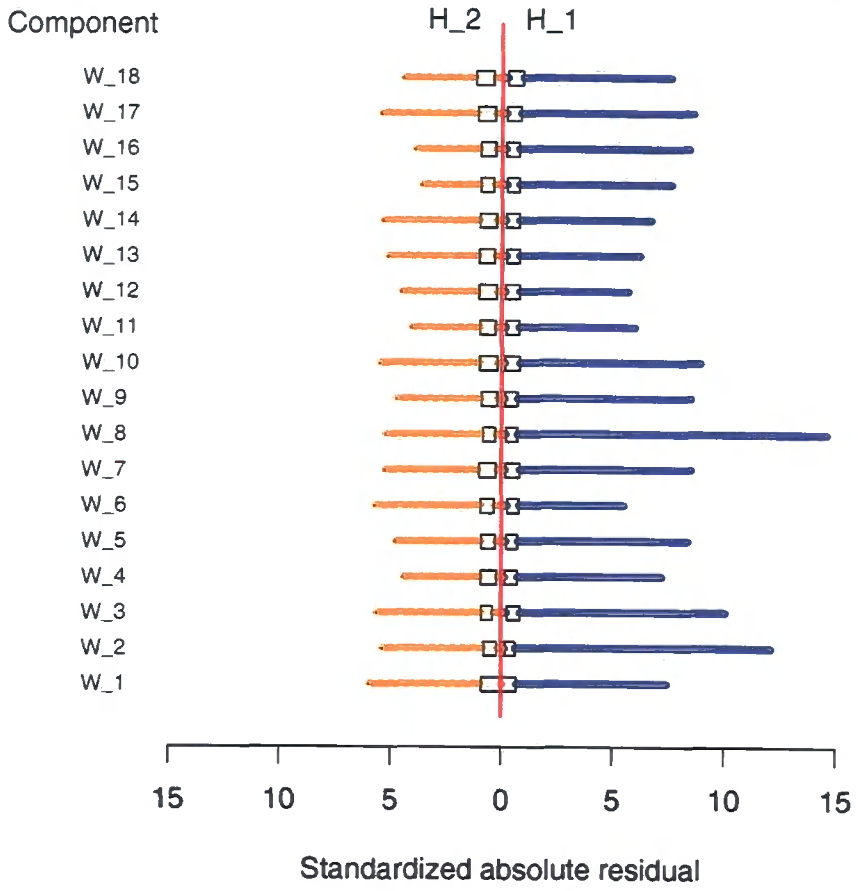


Figure 10.8: Distributions of the posttest residuals by components.

10.8 Comparison in the Equicorrelation Case

The variance comparison can be determined analytically for some cases. One case which might be of interest is the comparison when one model has homoscedastic independent components, and a second model has equally correlated components with the same variance. The comparison can be constructed as follows:

Consider two $k \times k$ matrices as follows:

$$\Sigma_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}, \qquad \Sigma_j = \sigma^2 I_k \tag{10.37}$$

Solving the eigenstructure problem of matrix $K = \Sigma_j^{-1}\Sigma_i$:

$$\begin{aligned} \text{Solve} \qquad \Sigma_j^{-1}\Sigma_i z &= \lambda z = \lambda I z \\ (\Sigma_j^{-1}\Sigma_i - \lambda I)z &= 0 \\ \det(\Sigma_j^{-1}\Sigma_i - \lambda I) &= 0 \end{aligned}$$

$\Sigma_j^{-1} = \frac{1}{\sigma^2} I_k$, therefore:

$$\Sigma_j^{-1}\Sigma_i = \begin{bmatrix} \frac{1}{\sigma^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma^2} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\sigma^2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{\sigma^2} \end{bmatrix} \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho & \dots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho & \sigma^2\rho & \sigma^2\rho & \dots & \sigma^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}.$$

Following that:

$$(\Sigma_j^{-1}\Sigma_i - \lambda I) = \begin{bmatrix} 1-\lambda & \rho & \rho & \dots & \rho \\ \rho & 1-\lambda & \rho & \dots & \rho \\ \rho & \rho & 1-\lambda & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1-\lambda \end{bmatrix} = \mathbf{C}$$

Solving the eigenvalue problem of $K = \Sigma_j^{-1}\Sigma_i$ is solving $|\mathbf{C}| = 0$. Following Graybill [37], the determinant of $\mathbf{C}$ is found by the following operations:

1. (i) Subtracting the original second row from the original first row to produce new first row; (ii) subtracting the original third row from the original second row to produce new second row; and so on. The result is the following:

$$\mathbf{C}* = \begin{bmatrix} (1-\lambda)-\rho & \rho-(1-\lambda) & 0 & 0 & \dots & 0 \\ 0 & (1-\lambda)-\rho & \rho-(1-\lambda) & 0 & \dots & 0 \\ 0 & 0 & (1-\lambda)-\rho & \rho-(1-\lambda) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \rho & \dots & (1-\lambda) \end{bmatrix}$$

2. Maintaining the first column; (i) the first column is added to the second column; (ii) the resulting second column is added to the third column; (iii) the resulting third column is added to the fourth column; and so on. The result is:

$$\mathbf{C}** = \begin{bmatrix} (1-\lambda)-\rho & 0 & 0 & \dots & 0 \\ 0 & (1-\lambda)-\rho & 0 & \dots & 0 \\ 0 & 0 & (1-\lambda)-\rho & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & 2\rho & 3\rho & \dots & (1-\lambda)+(k-\rho) \end{bmatrix}$$

3. Since $\mathbf{C}**$ is now a lower triangular matrix, its determinant is the product of its diagonals. The operations have not changed the determinant of matrix $\mathbf{C}$,

therefore:

$$|\mathrm{C}| = |\mathrm{C}^*| = |\mathrm{C}^{**}| = ((1-\lambda)-\rho)^{k-1}((1-\lambda)+(k-1)\rho)$$

4. Solving det $\mathbf{C} = 0$ gives us $\lambda = 1 - \rho$ for the first (k - 1) eigenvalues and $1 + (k-1)\rho$ for the last eigenvalue.

The first (k - 1) eigenvalues of matrix $K = \Sigma_j^{-1}\Sigma_i$ will be:

$$\lambda_1, ..., \lambda_{k-1} = 1 - \rho \tag{10.38}$$

and the last eigenvalue will be:

$$\lambda_k = 1 + (k-1)\rho \tag{10.39}$$

The first (k - 1) eigenvalues are always less than 1, and the last eigenvalue is always greater than 1, depending on the value of ρ. If relationship in (10.37) is reversed, solving the eigenstructure problem of $K = \Sigma_i^{-1}\Sigma_j$ will produce eigenvalues which are the reciprocals of (10.38) and (10.39), and the order reversed. Thus we have:

$$\lambda_1 = \frac{1}{1+(k-1)\rho} \tag{10.40}$$

$$\lambda_2, ..., \lambda_k = \frac{1}{1-\rho} \tag{10.41}$$

In this case, the first eigenvalue is always less than 1 and the following (k - 1) eigenvalues are always greater than 1, depending on the value of ρ.

10.8.1 Estimation of ρ with the Pretest Data

The condition that λ_1 is always less than 1 and the following $\lambda_2, ..., \lambda_k$ are always greater than 1 is representative of the outcomes of the pretest and the posttest data, where the majority of the eigenvalues are greater than 1 (The pretest eigenvalues are again presented in Table 10.7). Thus, it is plausible that the relationship proposed as (10.37) is a general representation of the covariance structure of the Learning data.

Table 10.7: Eigenvalues of compound matrix of pretest data.

W_j	*Eigenvalue*	W_j	*Eigenvalue*	W_j	*Eigenvalue*
1	3.35	7	2.17	13	1.63
2	3.14	8	2.01	14	1.56
3	2.63	9	1.92	15	1.27
4	2.56	10	1.73	16	0.91
5	2.24	11	1.68	17	0.38
6	2.19	12	1.66	18	0.15

Using the pretest data for illustration, the eigenvalues $\lambda_2, ..., \lambda_k$ which are greater than 1 (10.41) relate to the first $(k-1) = 17$ eigenvalues, and λ_1 which is less than 1 (10.40) relates to the last eigenvalue. Therefore from (10.40 and 10.41):

$$\rho = \frac{\frac{1}{\lambda_{18}} - 1}{(k-1)} \quad \text{and}$$

$$\rho = 1 - \frac{1}{\lambda_1} = ... = 1 - \frac{1}{\lambda_{17}}.$$

An estimate of ρ can then be computed by taking the average of all 18 estimates, as follows:

$$\begin{aligned}\hat{\rho} &= \frac{1}{k}\left(\frac{\frac{1}{\lambda_{18}} - 1}{(k-1)} + 1 - \frac{1}{\lambda_1} + 1 - \frac{1}{\lambda_2} + ... + 1 - \frac{1}{\lambda_{17}}\right) \\ &= 0.3358.\end{aligned}$$

This value of $\hat{\rho}$ is used to calculate the expected eigenvalues, had the covariance structure of the data been like (10.37). From Equations (10.40) and (10.41):

$$\hat{\lambda}_1 = \frac{1}{1 + (17)(0.3358)} = 0.1491 \qquad (10.42)$$

$$\hat{\lambda}_2 = ... = \hat{\lambda}_{18} = \frac{1}{1 - 0.3358} = 1.5056 \qquad (10.43)$$

The observed eigenvalues of the pretest data are plotted together with these expected values (10.42 and 10.43) in Figure 10.9.

There are three horizontal lines in Figure 10.9. The line in the middle shows 1.0, the value at which an eigenvalue would indicate equal variance under both

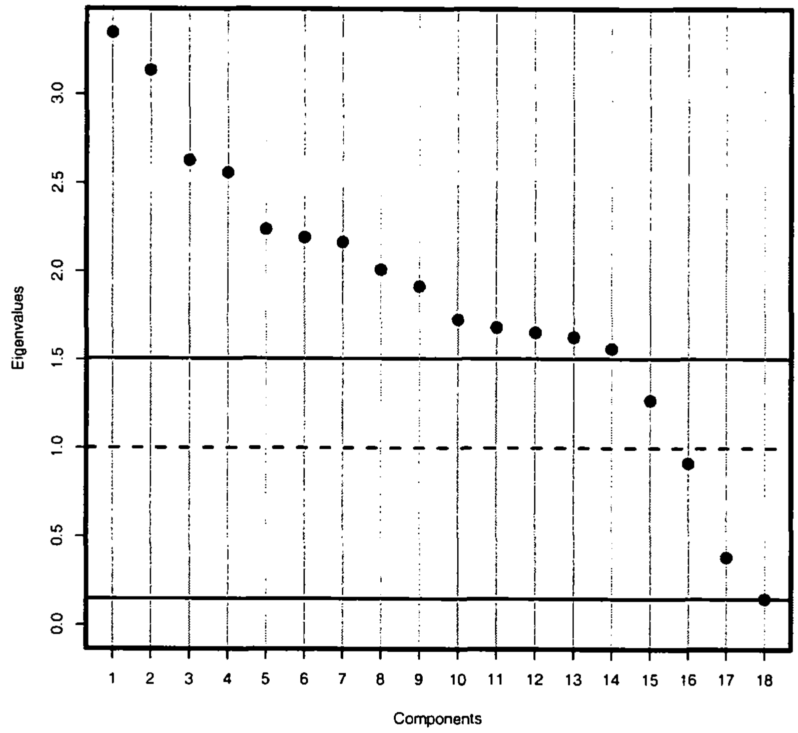


Figure 10.9: The eigenvalues and the lines of the expected eigenvalues.

hypotheses. The solid line at the bottom of the figure refers to the expected value of lambda (10.42) for all $\lambda < 1$, which are below the middle line. The top solid line refers to the expected value of lambda (10.43) for all $\lambda > 1$. The vertical distance between each eigenvalue to the respective line indicates the discrepancy for that particular component, between the observed covariance structure and the general structure as proposed in (10.37). The top line seems to be not far from the middle of all eigenvalues which are greater than 1, but the same thing cannot be said about the bottom line with regards to the eigenvalues which are less than 1. It lies near the last eigenvalue, and not in the middle of the three.

If the observed structure of the pretest data closely resembles the proposed relationship (10.37), we would expect the two solid lines to be positioned close to the middle of the distributions of their respective groups of eigenvalues. In the case of the Learning data, we may conclude that the structure within the Learning data is not like (10.37) very much. Nevertheless, it is not very far off either. We could test explicitly that the covariance structure is of a certain kind using tests of sphericity. See for example Krzanowski [57] page 166, for a single variance matrix.

10.9 Discrepancies Between Σ_1 and Σ_2

In Sections 10.6 and 10.7, we applied the method of variance model comparison to the pretest and the posttest data respectively. Each dataset produces its own eigenvalues (λ_j, $j = 1$ to k), which can be divided into two groups. The first group consists of those eigenvalues which are greater than 1, and the second group consists of those less than 1. Eigenvalues in the first group represent components that have greater variance under hypothesis 1 than they do under hypothesis 2. Eigenvalues in the second group represent components that have lower variance under hypothesis 1 than under hypothesis 2. λ_1, which is in group 1, and λ_{18}, which is in group 2, represent the components with the largest discrepancy between hypothesis 1 and hypothesis 2, in each group respectively. Comparing the eigenvalues from the different datasets could provide some information on how the discrepancies are, whether they are more or less the same in both the pretest and the posttest data, or whether they are totally different.

Eigenvalues of the pretest and the posttest data are plotted in Figures 10.1 and 10.5 on pages 274 and 281 respectively. Values of τ are calculated along with e^τ, to indicate the amount of discrepancies between hypothesis 1 and hypothesis 2 within the datasets. The statistics of the two datafiles are presented in the following table:

Statistics of eigenvalues from both datasets.

	Pretest	*Posttest*
$\Sigma_{j=1}^{k} \mid log\lambda_j \mid$	13.66	14.11
$\tau = \frac{1}{k}\Sigma_{j=1}^{k} \mid log\lambda_j \mid$	0.76	0.78
e^τ	2.14	2.19
λ_1	3.35	3.33
λ_{18}	0.15	0.13

Values in the first three rows indicate that the posttest data have slightly more discrepancies between hypothesis 1 and hypothesis 2, than do the pretest data. The values of λ_1 and λ_{18} of the pretest are just slightly larger than those of the posttest data. This suggest that the largest discrepancies among the components with $\lambda > 1$ and among the components with $\lambda < 1$ occur within the pretest data.

Chapter 11

Conclusions and Discussions

This study was initiated by the need of the National Institute of Public Administration (INTAN) to measure its performance. Being a training institute that serves the majority of the Malaysian public sector officers, INTAN should have its performance assessed mainly by the *effectiveness* of the training programmes it runs, and not simply by a measure that indicates *customers' satisfaction.* We take *effectiveness* to mean impact on the participants with regards to the three aspects of training; *attitude, skills*, and *knowledge.* Without these impacts being measured, there is no indication for the effectiveness of the training programmes.

The number of training programmes per year at INTAN has always been on the increase, with the latest figure at over 1,300 programmes in 2006 [40]. So does the number of participants increase, with the latest statistics indicating close to 46,000 participants attending training programmes at INTAN in the same year. With the recent emphasis by the Malaysian Government on the contribution of INTAN to the efficiency of the public service delivery system, the impact of training on the participants needs to be even more clear [62].

The current programme evaluation is designed to indicate mainly the satisfaction of the participants on three aspects of the training programmes, namely (i) overall management, (ii) techniques of training used, and (iii) contents of the programmes; and their perception of (i) achievement of objectives, (ii) effectiveness of programmes, and (iii) benefits of the course (Please refer to Figure 1.1 on page 5). Even though *effectiveness of programmes* is in the model, it is indicated by just two questions, each asking the participants whether they think there has been an increase in their *skills* or *knowledge*. Thus, *effectiveness* of the programme is not measured by a properly structured questionnaire, but relies on the participants' immediate reaction.

This study suggests a new approach of evaluation, which makes use of three questionnaires, namely the GHQ, the CEQ, and the LQ. We develop the LQ based on a Learning Model and evaluate its value, and we evaluate the possibility of using the other two questionnaires. We discuss and recommend the use of the three questionnaires, as tools in evaluating the effectiveness of the training programmes. We also propose a method of graphical comparisons of structural equation models which is based on Goldstein and Wooff [36], as well as a new index for evaluating the reliability of scales. Another way of analytically determining the variance structure (for some cases) is also discussed.

All the work that this study encompasses is presented in Figure 11.1 on page 297. The figure is colour-coded; blue indicates general work, bright green is for work on the GHQ, red is for work on the LQ, light brown indicates work on the CEQ, and darker green is for work on the variance models comparisons. Yellow boxes indicate the three main steps of the study. Boxes with shadows indicate the main products of the study.

Discussions in this chapter are divided into six sections. The first two sections are devoted to the statistical contributions of this thesis. In Section 11.1 are the two main original statistical contributions, which are the methodology for graphical comparison of structural equation models, and the proposed index of alpha*, as one way of examining the reliability of scales. In Section 11.2 we summarize the use of the questionnaires to help solve the problem of training evaluation at INTAN.

In Sections 11.3, 11.4, and 11.5 are discussions on the three questionnaires: the LQ, the GHQ and the CEQ respectively. The LQ is the newly developed questionnaire, designed to measure 'learning'. The proposed Learning Model has been shown to be not consistent with the data and requires modifications. Nevertheless, the LQ does detect pretest-posttest changes in the scores of two of its subscales, which indicates that the measure is sensitive to the changes. The measurement model of

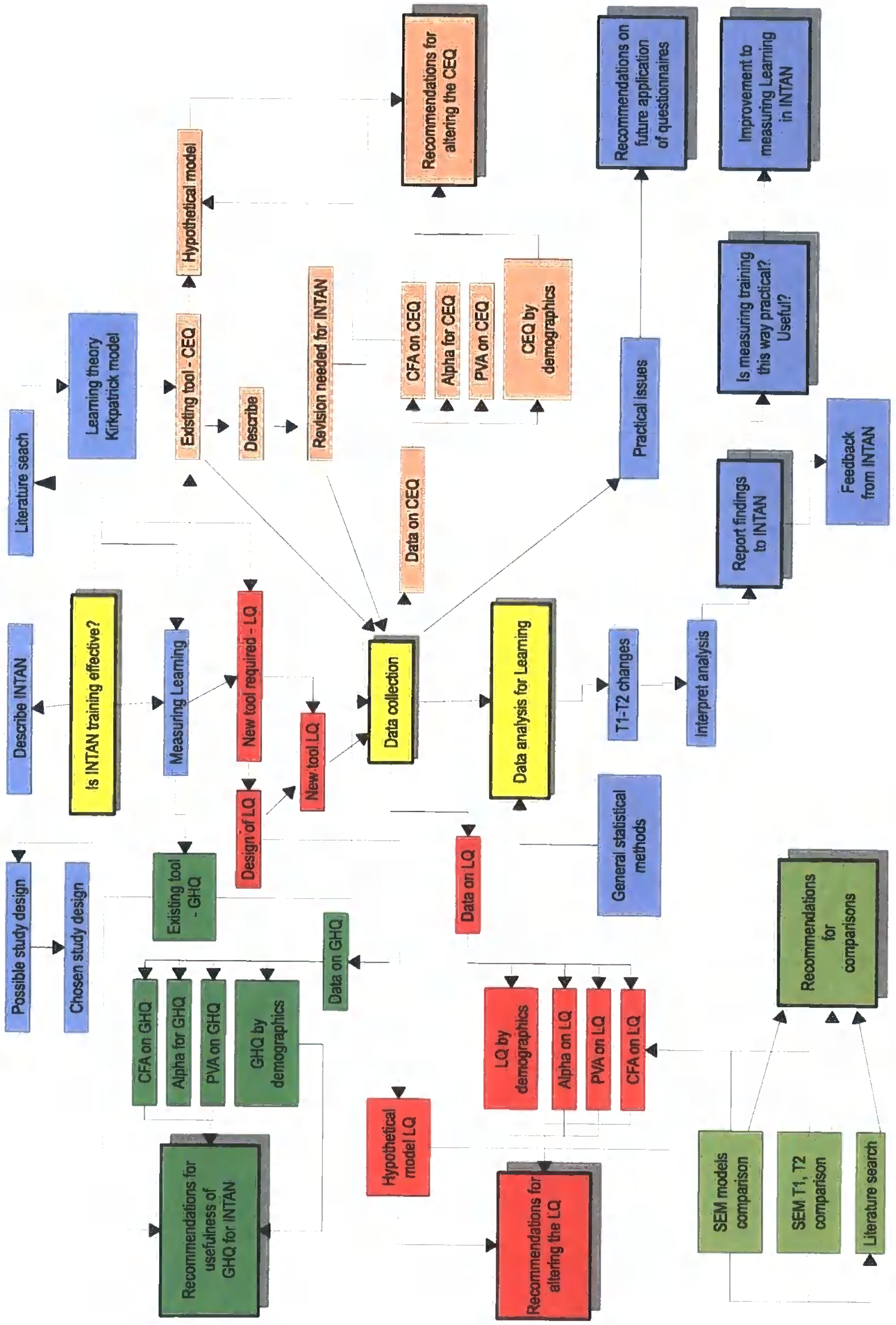


Figure 11.1: Overall plan of the study.

the second questionnaire tested, the GHQ, has also been shown to be not very consistent with the data. However it appears to have one single dominant factor, and this agrees with the proposed model. The CEQ has five different subscales. The measurement models of four of the subscales fit the data well. The fifth subscale which has three items does not seem to fit the data very well.

Section 11.6 is confined to the discussion about the contribution of this thesis to INTAN as a training institute. The original problem is the attempt to measure 'learning', which would indicate training effectiveness better than the current evaluation method. This study recommends the use of the CEQ to measure *reaction*, and the modified version of the LQ to measure *learning*, as well as the GHQ to indicate the general psychological health of the participants.

11.1 Statistical Contribution

There are two main statistical contributions of this thesis; the graphical comparison of structural equation models, and the index of alpha*. The former is discussed in Chapter 10, while the latter is used in Chapter 5.

11.1.1 Graphical comparison of SEM models

In Chapter 10, which starts on page 259, we present the graphical method of variance model comparison. The method is an adaptation of Goldstein and Wooff [36], where variance matrices estimated given two differing SEM models are compared. In this thesis, the method is applied to the Learning data, based on the modified Learning Model (Section 6.3.2 on page 109). In this modified LQ, there are eighteen items instead of thirty as in the originally proposed Learning Model.

The two SEM models are given by the hypothesized structures of the variance matrices of the data, dubbed as Σ_1 and Σ_2. The first represents the hypothesis that the eighteen indicators of Learning are independent of each other. Σ_1 is thus diagonal, with the diagonal entries estimated from the sample. The second is the model-implied covariance matrix, which estimates the population model-implied structure. The model is based on the new Learning Model, as mentioned above.

The model-implied covariance matrix is produced by the SEM analysis using the maximum likelihood minimizing function.

How the method works

This method tests the hypotheses presented in Section 10.3 on page 262. How the eigenstructure of the matrix $K = \Sigma_2^{-1}\Sigma_1$ is solved is also presented in detail in that section. To examine which of the two hypotheses fit the data better, discrepancies between Σ_1 and Σ_2 are examined by several ways: (i) graphical comparisons using several plots, (ii) looking at the eigenvalues of the K matrix, and (iii) calculating the indices of *tau* (τ) $= \frac{1}{k}\Sigma_{i=1}^{k}|log\lambda_i|$, and e^{τ}.

Plots : Two informal plots are used to examine the discrepancies between the two hypotheses, namely the plot of the negative log(p-values) of the K-S test (Figures 10.3 on page 279 and 10.7 on page 286), and the distributions of the standardized absolute residuals (Figures 10.4 on page 280 and 10.8 on page 287). The first plot reflects the probability of rejecting the hypothesis that the residuals are Normally distributed with mean zero and variance one ($N(0,1)$). The plots for both the pretest and the posttest data suggest that neither models seem to generate $N(0,1)$ residuals, but model 1 seems more abnormal. The second plot is a presentation of the boxplots of the absolute values of the residuals. In both the pretest and the posttest data, the distributions of the residuals under H_1 are generally wider than those under H_2, suggesting that the data fits the second hypothesis better that it does the first one.

The eigenvalues : As explained in Section 10.3.1 on page 267, each of the eigenvalues $\Lambda = \lambda_i, ..., \lambda_k$ indicates whether component W_j has larger variance under H_1 or H_2. Plots in Figure 10.1 on page 274, and Figure 10.5 on page 281 present a simple graphical way of examining the eigenvalues. Both plots indicate the same findings; in both the pretest and the posttest data, there are more components that have larger variance under H_1 than those that have larger variance under H_2.

Tau : τ is an index which indicates the amount of discrepancies between the two

> hypotheses. If all components W_j have equal variance under both hypotheses, τ would equal to zero, and e^τ would equal to one. The further τ is from zero, or e^τ is from one, the larger the discrepancies are between Σ_1 and Σ_2. As reported in Section 10.9 on page 293, the value of τ is 0.76 for the pretest data, and 0.78 for the posttest data. The equivalent values of e^τ are 2.14 and 2.19 respectively. These indicate that the discrepancies between the two hypotheses are just slightly larger with the posttest data.

We also look at the possibility of analytically determining the variance comparison for some cases, specifically when one model has homoscedastic independent components and the other model has equally correlated components with equal variance. The discussion is presented in Section 10.8 on page 288. In this discussion we suggest that the eigenvalues could be estimated, and then the estimated eigenvalues are compared with the observed values. If the actual structures are close to what is suggested, the expected eigenvalues would fall somewhere in the middle of their respective group's distribution. This method is applied to the pretest data and the finding is that the structure within the Learning data is not very close to what is suggested.

11.1.2 Alpha*

In Section 3.7.1 on page 49 we introduce an alternative measure which we call alpha*, to examine the reliability of the scales. While the Cronbach's coefficient alpha is the common measure of reliability, it has been shown that the value increases with the number of items [67]. Multiple-item scales could have high reliability because of high correlation between the items and the latent variable it is suppose to measure, or, it could have high reliability because it has many items which correlate weakly with the latent variable. The measure of alpha* gives the implied reliability had the scales had two items, thus it is suitable to be used in comparing the reliability of scales with different numbers of items.

In Chapter 5, which starts on page 75, the values of Cronbach's alpha and alpha* are calculated for all scales and subscales. A large difference between the two indices suggest that the scale has a lot of items to compensate for weak correlation between the items and the latent variable it is suppose to measure. Even though alpha* does not measure reliability directly, it is useful in comparing the reliability of different scales.

11.2 INTAN

Training evaluation at INTAN has always been done at the *reaction* level (Kirkpatrick's model of evaluation, Section 2.1 on page 11). For the management of INTAN to know how effective the training programmes really are, programme evaluation needs to be carried out at a 'higher' level. The current programme evaluation is not able to indicate *effectiveness*, which is defined as impact of the programmes on the three aspects of training; *attitude*, *skills*, and *knowledge*. Until we find a way to do this, the actual performance of INTAN is not clearly known. A literature search failed to find a suitable measure that can be used at INTAN, thus this study attempts to develop a questionnaire, which we call the LQ, to measure *learning*, which is at the second level of the Kirkpatrick's model.

In the Learning Model which is the basis of the LQ, the factor Learning is measured through three latent variables, namely knowledge, application, and importance. This is shown in Figure 4.1 on page 64. Each of the latent variables, which is a subscale of the LQ, is in turn indicated by ten items that represent ten subject areas. Details of the items are presented in Table 4.1 on page 68.

Findings of this study suggest that the LQ is capable of detecting changes between the pretest and the posttest scores of two subscales, namely the knowledge and the importance. The difference in findings between the study group and the control group supports the idea that the changes are due to the training attended by the study participants. On the other hand, we also find that the Learning Model needs modifications, where each item from the four highly correlated items are taken out from each subscale. The suggested modifications are summarized in Section 11.3.1 in this chapter. With the modifications, the new model needs to be further evaluated with empirical data.

Besides the LQ, this study also makes use of two other questionnaires: the GHQ to measure general *psychological health*, and the CEQ to measure *course experience*. We propose that the CEQ be used in INTAN, as another measure of *reaction* which would complement the current programmes evaluation. We also propose the use of the GHQ to provide background information on the general psychological health of the participants.

11.2.1 Usefulness of the questionaires in INTAN

The LQ This newly developed questionnaire is found to be useful as it is, but we propose further evaluation on the new model. The LQ provides an indication of the effectiveness of the training programmes on the participants, with respect to the three LQ subscales. A clear indication on the effeciveness of training is what is lacking in the current programme evaluation at INTAN. To evaluate the value of the new LQ, new empirical data is needed. We recommend the use of the modified LQ immediately, mainly to collect further data. The modifications of the LQ are as given in Section 11.3.1 below.

The GHQ There has never been an evaluation on the participants' general psychological health before in INTAN, but the extra information it gives should be able to be put to good use. A cross-factor study would be able to indicate, for example, whether the performance of the participants in training is associated with their psychological health. If it was found to be so, then psychological health would be one of the factors to be considered in designing training programmes. The GHQ may be used without any modifications.

The CEQ The CEQ provides another perspective on the 'reaction' of the training participants on the training programmes they attend. The findings will be a valuable complement to the findings of the current evaluation method, as the CEQ measures five different aspects of training, most of which are not covered by the current programme evaluation. There is no modification necessary, but conclusions on the appropriate assessment subscale should be made with caution.

11.3 The Learning Questionnaire

The LQ is specially developed to measure *learning*, a concept closely related to the second level of the Kirkpatrick's model of evaluation [52]. We have introduced the background problem in Section 1.2 on page 3, and discussed in details in Chapter 4 why this kind of measure is needed in INTAN.

The samples on which the LQ is tested and evaluated consist of 760 training participants from the National Institute of Public Administration (INTAN), the main training institute for the Malaysian Public Service. The samples represent two main service groups of the Malaysian Public Service, namely the Professional and Administrative Officers group and the Supporting Staff group 1. Two types of sample are used in this study, the treatment group and the control group, with the only difference between them being whether they attend training at INTAN at the time of study or not.

11.3.1 Summary findings

Reliability

The main scale of both the pretest and the posttest LQ seem to be reliable, with the values of the Cronbach's alphas being greater than 0.85, as indicated in Chapter 5 which starts on page 75. All subscales also have alpha values greater than 0.85. For both the pretest and the posttest scores, the importance subscale has the highest alpha, followed by the application subscale, then by the knowledge subscale. The percentage of reduction from Cronbach's alpha to alpha* varies among the subscales, implying differences in the amount of dependency on the number of items in the scales. The knowledge subscale has the highest reduction, followed by the application, then by the importance subscale. This result is similar in both the pretest and the posttest data. These seem to suggest that among the three LQ subscales, importance is the most reliable, and knowledge is the least.

CFA, EFA and PVA

The main drawback in the development of the LQ is the fact that none of its measurement models are consistent with the data. None of the fit indices show values anywhere near the good-fit thresholds. Modification indices produced by the SEM software suggest four highly correlated pairs of indicator items across the three subscales, namely the pairs of items 1 and 2, items 5 and 6, items 7 and 8, and items 9 and 10.

The exploratory factor analysis shows that there might be only two dominant factors in the structural model instead of three as proposed. The good news here is that all items that are suppose to indicate knowledge load on one factor, while all other indicator items load on the other factor. The contribution of each items to the overall variance is also highly variable as indicated by the principal variable analysis (Chapter 7, starts on page 117). Almost 90% of the total variation in each subscale could be explained by just six items of that subscale, and this suggests strongly the need for dimension reduction.

For further work on the development of the LQ scale, we suggest a new Learning Model after considering all the above results. We also present two approaches to selecting the indicator items: one is based on the Cronbach's alpha values, and the second one is based on the PVA results. The new Learning`Model is developed by taking out one item from each of the highly correlated pairs of items. Using the first approach, the item that reduces Cronbach's coefficient alpha the most is taken out. As a result, each subscale consists only six items as in the following table. The details of the items are presented in Table 4.1 on page 68.

New Learning Model based on alpha.

Subscale	Items taken out	Items left in model
Knowledge	2, 6, 8, 9	1, 3, 4, 5, 7, 10
Application	2, 6, 7, 10	1, 3, 4, 5, 8, 9
Importance	2, 6, 7, 10	1, 2, 3, 4, 8, 9

If the principal variables analysis is used as the basis for choosing the items to be taken out, the result would be a little different. Between items 1 and 2 in the knowledge subscale, item 2 has been shown to contain more information than item

1, thus the former should be retained. The same goes with the pair of items 9 and 10, where item 9 appears to contain more information. For the correlated items of the application subscale, item 1 is a better candidate to be taken out than item 2. Similarly item 9 is better taken out than item 10. In the subscale of importance, item 2 should be retained instead of item 1, item 7 should be retained instead of item 8, and likewise, item 10 instead of item 9. Therefore based on the principal variables analysis, a new Learning model would consist of indicator items as in the following table.

New Learning Model based on PVA.

Subscales	Items taken out	Items left
Knowledge	1, 6, 8, 10	2, 3, 4, 5, 7, 9
Application	1, 6, 8, 9	2, 3, 4, 5, 7, 10
Importance	1, 6, 8, 9	2, 3, 4, 5, 7, 10

As the contribution of each item to the overall variation is more important than a small change in the Cronbach's alpha, we propose that the modified LQ is based on the findings of the PVA, as explained above.

Tests of differences

The LQ fails to indicate any differences between or among the levels of the seven demographic factors. This is explained in Chapter 8 which starts on page 129. We thought of two possible reasons for this: one is that the LQ is not sensitive enough to the differences, and this could be because of the small sample sizes in some levels. The second possible reason is that there are no genuine differences between or among the demographic levels. However, one interesting finding is that when the LQ is used to compare the scores of the pretest and the posttest (Chapter 9), the differences in the scores of two subscales are detected and shown to be statistically significant. This suggests that the LQ is sensitive to detect differences in the scores of the LQ subscales, so the failure to do so between and among the factor levels is likely to be due to the small sample sizes.

ANCOVA

The ANCOVA in Section 9.4 on page 209 is the better method for examining the association between the demographic factors and the posttest scores of the LQ subscales. Of the seven demographic factors, only *experience*, *centre* and *course* seem to show some level of association with the subscales. It is likely that the factor of posttest *experience* is associated with application and importance subscales, after adjusting for the pretest score. There is also slight evidence that the posttest knowledge and application are associated with *centre*, after adjusting for the prestest scores. The factor of *course* appears to be associated with all three LQ subscales. It suggests that changes in Learning occur differently among the different courses.

11.4 The GHQ

The GHQ is a unidimensional instrument, designed to screen common mental disorder [93], to identify and measure psychological problem [18], or to detect non-psychotic psychiatric disorder [75]. In this study the instrument is used on normal and healthy respondents, thus the score is related to general psychological health. The factor of general health is indicated by twelve items, but the structure has been shown to be not very stable [93]. Results of analysis in this study indicate similar findings; the model does not fit the data very well.

11.4.1 Summary findings from this study

Reliability

The GHQ is hypothesized to be a one-dimensional measure, thus there is only the measurement model to evaluate. Applied to the data in this study, the Cronbach's alpha value of the GHQ is 0.8845. The value of the alpha* is 0.5606 which may be considered to be relatively high.

CFA

Discussions in Section 6.2.1 on page 98 suggest that the measurement model of the GHQ does fit the data, but the fit is not very good. Only one of the fit indices (The RMR) shows a value not too far off from the good-fit zone. Because the fit is not good, modification indices are produced. There appears to be two pairs involving three indicator items, which are correlated and may contribute to the model being not consistent with the data. The pairs of indicators are items 3 and 4, and items 4 and 8. We suspect that the respondents found it difficult to differentiate between these statements.

In spite of the poor fit, the scree plot indicates that there is one dominant factor in the GHQ structure, and this is in agreement with the original model. The exploratory factor analysis does suggest three different factors, but the items that load on the second and third factor also load on the first factor, thus the extra factors may not be read into too much.

PVA

Principal variables analysis on the GHQ scale (Section 7.3 on page 126) suggests that the amount of information contributed by each of the twelve items do not differ too much. Item 9 (*Been feeling unhappy and depressed*?) appears to contribute about 29% to the overall variation, followed by item 4 (*Felt capable of making decisions about things?*) with about 15% contribution. With these two items in the scale, the contribution of the rest seem to be about equal. Among the three highly correlated items (items 3, 4, and 8), item 4 is clearly the most valuable to the scale. The other two are not as valuable, with item 3 contributing to about just 5% and item 8 to about 4%.

Tests of differences

The score of the GHQ seems to be associated with five of the seven demographic factors, ie. all except *sex* and *ethnicity*. Otherwise, the score of the GHQ seems to differ between or among the different levels of *age, centre, service sector, service group*, and *experience*. This suggests that general psychological health differs among respondents from the different age groups, the different centres, the two service sectors, the different service groups and the different experience groups. However, the score of the GHQ does not seem to be associated with the changes in the three LQ subscales.

11.5 The CEQ

The CEQ is widely used in the Australian higher education system, and is starting to be used in some institutions in the UK [96]. It was designed to measure differences in the quality of teaching between comparable units in the higher education system [24]. The details of the development work on the CEQ are discussed in Section 3.4.1 on page 40. In this study, the instrument is evaluated as a candidate questionnaire to measure the *reaction* of participants to the training programmes at INTAN. It would be a valuable complement to the current programme evaluation.

11.5.1 Summary findings from this study

Reliability

The CEQ is most commonly not measured as a factor, but consists of five different subscales, each measuring different aspects of training. The reliability of the subscales varies as indicated by the Cronbach's alpha values in Table 5.1 on page 76. The GT subscale seems to be the most reliable with alpha of 0.8730, while the AA subscale seems to be the least reliable with alpha of 0.4866. The AA subscale also shows the smallest alpha* value of 0.3872, suggesting low reliability. However, the largest reduction from alpha to alpha* is shown by the CG subscale with slightly over 26%, and the AW subscale shows the least reduction of 18.43%.

CFA

The confirmatory factor analysis via the structural equation modeling on the measurement models indicate good fit for four subscales, as discussed in Section 6.1 on page 89. One subscale, the AA, shows poor fit, which could be due to the fact that it has only three indicators.

The CG subscale seems to be most valid, with almost all fit indices showing values of good fit. They include the χ^2, the CMIN/DF, the RMR, the GFI, the AGFI, the NFI, the CFI, and the RMSEA. The subscales of GT, GS and AW also perform well, with five of the fit indices indicate good-fit, namely the RMR, the GFI, the AGFI, the NFI, and the CFI. The AA subscale only has the indices of GFI, and maybe the AGFI, that show values anywhere near good-fit thresholds.

Two structural models are tested: the standard CFA model (Figure 6.1 on page 90) and the two-factor model (Figure 6.2 on page 95) which was proposed by Wilson et al. [96]. The fits of both of the models are not very good, suggesting that they are not very consistent with the data. Fit indices for both models indicate varied results, with only the GFI and the AGFI showing values not too far from the good-fit thresholds.

PVA

Each of the CEQ subscales consists of between three to six items, thus the PVA is not very appropriate at the subscales' levels. The PVA is carried out on the main scale, where all items of the subscales are combined together and the contribution of each item to the overall variation of the main scale is examined. The results of the analysis as presented in Section 7.4 on page 127 suggest that the item that contributes the most is item 10, which is an item of the GS subscale. It contributes to about 21% of the total variation. One observation from this analysis is that the first five items that contribute the most variation come from five different subscales.

Tests of differences

The scores of all five of the CEQ subscales do not differ between male and female respondents, nor among the ethnic groups. With the other demographic factors the results vary, as summarized in Section 8.10 on page 188. The scores of the GT and the GS subscale shows strong evidence of differences between or among the levels of *age, centre, service sector, service group*, and *experience*. The score of the CG indicate strong evidence for differences among the *centres* and among the *service groups*, and weaker evidence for *age* and *service sector*. There is only some evidence of differences in the score of the AA subscale among the *centres* and *service groups*. There is a strong evidence for differences in the score of the AW among the *centres* and the *service groups*, and weaker evidence likewise for *service sector* and *experience*.

11.6 Discussions for INTAN

As the main training institute for the Malaysian Public Service, INTAN needs to know how effective its training programmes are. An effective training programme has impact on the participants in terms of their *knowledge*, their *skills*, and their *attitude*. In the current programme evaluation model, the impact of training programmes on these three aspects of training among the participants is not clearly measured. Consequently, there is no proper indication of the effectiveness of the training programmes.

This study does three things: (i) develop the LQ which attempts to measure *learning*, a factor positioned at the second level of the Kirkpatrick's four levels of evaluation model (The model is explained in Section 2.1 on page 11). This factor relates directly to the three aspects of learning (*knowledge, skills*, and *attitude*), thus the measurement would indicate effectiveness of training more clearly than the currently used programme evaluation; (ii) evaluate the usefulness of the CEQ and the GHQ questionnaires on the training evaluation at INTAN. The CEQ measures *course experience*, which means it would make a valuable complement to the current evaluation in examining participants' satisfaction. The GHQ measures general psychological health, and it could be useful to the overall evaluation, though not as directly as the other two questionnaires; and (iii) examine the associations between the scores of the questionnaires and the seven demographic factors. Some findings are more interesting than the others, such as the association between the course experience and *centre*. The results of the tests are discussed in Chapter 8 on page 129.

11.6.1 Measuring 'reaction' and 'learning'

In this study, the LQ has been shown to be sensitive to the differences between the pretest and the posttest scores of two LQ subscales, namely the knowledge and the importance. If the factor of Learning is to be measured among the training participants, it would take the pretest and the posttest approach just like in this study. Therefore, the findings of the pretest-posttest differences among the study samples is supportive of the idea that the factor of Learning can be measured by the LQ instrument. With the LQ, the main question that would be answered is:

> Are there any changes in the level of knowledge, the level of application and use, or the level of importance during training?

The three factors underlined above indicate Learning, thus any changes in the scores of the factors over the training period would suggest a change in Learning as an impact of the training. However, this study has also indicated the need for modifying the Learning Model on which the LQ is based. As the new model has to be evaluated with empirical data, we propose that a new set of data is collected based on the modified LQ, which is explained in Section 11.3.1 above.

With regards to the CEQ, this study has shown that at least four of its five subscales are valid with this sample, as discussed in Section 6.1 on page 89. The appropriate assessment subscale has some problem with its validity, but the drawback is minor and does not require it to be taken out altogether. The five aspects of training the CEQ measures (good teaching, clear goals, generic skills, appropriate assessment, and appropriate workload) are all very relevant to INTAN's interest, and the findings would be valuable complements to the current programme evaluation. With the findings of the CEQ, the following questions would be answered:

1. Do the facilitators (trainers) appear to practise good teaching habits?
2. Are the participants clear about what is expected of them in the training?
3. Do the participants see the training programmes as providing the necessary skills?

4. Do the participants think they are assessed appropriately?

5. What do the participants feel about the work pressure during the training?

In INTAN's current programme evaluation (The model of which is presented in Figure 1.1 on page 5), three aspects of facilitators (trainers) are assessed, namely (i) openness, (ii) responsiveness, and (iii) general management. An answer to question 1 would complement those aspects, by specifically targetting what the participants would view as good teaching practices shown by the trainers. The other four aspects measured by the CEQ as mentioned in the above paragraph are not in the currently used programme evaluation, thus they would all be valuable complements, from which the findings would give clearer indication of the participants' satisfaction.

In Section 9.4.3 on page 212, we examine the association between several demographic factors, and the scores of the LQ subscales. Among the factors is *course*, which refers to the individual training programmes. We find that T2-T1 changes in the scores differ among the different courses, suggesting that the participants 'learn' differently. Figures 9.11, 9.19 and 9.27 on pages 223, 237 and 248 respectively, are examples of an easy way of checking the performance of the individual courses. Ideally, we would like to see each course shows a distribution above the zero line, which would indicate an increase in the score from before a course starts, to the time it ends. A course that indicates decreases in all three subscales for example, might imply an underlying problem that needs attention. It has also been suggested in Section 9.4.8 that participants who come to training while being stressed tend to get less positive training experience compared to those who are not stressed.

11.6.2 Distribution of the questionnaires

The LQ is divided into two sets: the pretest and the posttest. The pretest has to be administered before training starts, and the posttest after training finishes. The CEQ has to be completed upon completion of the training programmes. The GHQ can be administered any time; thus before training starts might be a better option, as we do not want to take too much of the participants' time after the programmes end.

All questionnaires should be incorporated into the electronic evaluation system, where participants respond only through the network-linked computers. This saves time, as those questionnaires administered before training starts can be done together with the registration, which all respondents are required to do. The after training questionnaires, ie. the CEQ and the posttest LQ, can be administered together with the normal programme evaluation. The extra time needed in each session should be not more than about ten minutes, which is quite negligible.

11.7 Discussion for general training evaluation

Evaluation of training programmes at the level of *reaction*[1] is probably the easiest to do. However, in terms of the actual impact of training on the participants, evaluation carried out at this level is least useful. It does not indicate any changes in the three aspects of learning; *attitude, skills* and *knowledge.* Unless there is some indication of changes in these aspects between before the programme starts to the time it ends, the impact of training is not clearly known. In other words, we do not have a clear idea whether the training programmes have been effective or not. Thus, in any training institute, evaluation of training programmes is almost always better to be carried out at a level higher than the *reaction.* In the Kirkpatrick's model, the higher levels are *learning, behaviour*, and *results.* It would be wise to concentrate on the next higher level, ie. *learning*, rather than the other two which are known to be very hard to measure.

Evaluation at the level of *learning* would be easier to be carried out on programmes which are technical in nature, than on programmes which are naturally more subjective. Similarly, it would be easier to evaluate changes in knowledge of a specific subject, than changes in knowledge over many different subjects. For a training institute like INTAN, it would make more sense to have one single evaluation tool, than to have separate evaluation questionnaire for each programme. The tool then must be suitable to be used on all programmes, yet its sensitivity to changes in the training aspects it is supposed to measure, understandably, must not

[1] As in Kirkpatrick's level of evaluation

be compromised.

This study has proposed that the requirement for that kind of evaluation tool could be fulfilled by developing a specific evaluation model, which is based on the range of subjects taught at the institute. For each of the subjects, training participants are to give their perceived scores on *knowledge* and *skills*, and scores on the importance of learning the subject, which indicate their *attitude* towards learning. To be able to make more meaningful conclusions from the data, evaluation is done at two time points; first before the training programme starts, and the second one after it ends. Changes in the scores between these two time points are taken as indications of the effectiveness of the programmes. In this thesis, the specific evaluation model is translated into the Learning Questionnaire (LQ) which consists of ten indicator items repeated in three sections. Depending on the range of subjects taught at a different training institute, the indicator items vary.

Findings from this study have shown that not all of the subjects are equal in their contribution to the measurement of *learning*. This is especially explained in Sections 6.3 and 7.1. Based on the results of the analyses in these two sections, two new Learning Models are proposed; one based on the SEM and the other one based on Principal Variables. I have indicated that the latter is more preferable to the former. The fact that the indicator items contribute differently to the score of *learning* is also expected to be the case with a different set of subjects. If initial data show that the differences are significant like in the case of LQ in this thesis, the model should be modified and evaluated by a new set of data.

11.8 Overall conclusions

Much attention in this thesis is focused on the statistical analyses of the results from the three questionnaires, the LQ, the CEQ, and the GHQ. Chapters 5, 6, 7, 8, and 9 all discuss the findings of the surveys. However, one of the key purposes of this thesis is to tackle such datasets in generality, and this is presented in length in Chapter 10. Here we propose two main statistical contributions of the thesis, namely (i) the graphical comparison of structural equation models, and (ii) the index of alpha*. In (i), the main objective is to compare two competing hypotheses with the observed covariance structure, to see whether one of the hypotheses fits the data, or both hypotheses fit the data, or none of them fits the data. Related to that, we also calculate τ and e^{τ}, two indices that indicate the amount of discrepancies between the two hypotheses. In principle, tests could be devised to test for improbably large discrepancies, but this is outside the scope of this thesis. In (ii), the index shows the implied reliability for a scale with two items. It is appropriate to be used in examining the reliability of scales with different numbers of items.

The other key purpose of this thesis is the development of the LQ, which is a new questionnaire designed to measure 'learning', which we hope will improve the programme evaluation at INTAN. We have shown that the LQ is capable of measuring the pretest-posttest change in each of the LQ subscales. It still needs to be modified for improvement, and INTAN will provide the datasets needed for further evaluation. We have also shown that the CEQ and the GHQ are usable in INTAN without any modification. They will become complementary tools that will help improve programme evaluation. The CEQ evaluates five different aspects of training, while the GHQ provides for cross-factor analysis of training effectiveness.

Another part of the thesis is the results or findings of the three questionnaires with regards to the respondents of this study. One of the most important findings is that 'learning takes place differently among some of the different subgroups of the training participants. Besides 'learning', 'course experience' scores and 'general health' scores have also been shown to differ among some of the different subgroups. The differences among the subgroups are expected, but the framework for comparing the subgroups will be in place once the questionnaires are fully utilized as programme evaluation tools.

As an immediate future development of the LQ, the modified version should be used to collect a new set of empirical data from INTAN. The new data will be used to evaluate the new model, thus continuing the development of the 'learning' measurement tool. With regards to the variance model comparison, future work could look at the probability distribution for eigenvalues $\lambda_1, ..., \lambda_k$, and could also bootstrap the summary statistics.

Bibliography

[1] *Advanced Learner's Dictionary*. Oxford University Press, Oxford, 2000.

[2] K. J. Abernathy. Foraging ecology of Hawaiian monk seals at French Frigate Shoals, Hawaii. Master's thesis, University of Minnesota, 1999.

[3] G. M. Alliger and E. A. Janak. Kirkpatrick's levels of training criteria: thirty years later. *Personnel Psychology*, 42:331–342, 1989.

[4] J. L. Arbuckle and W. Wothke. *Amos 4.0 Users' Guide*. SmallWaters Corporation, Chicago, 1999.

[5] James L. Arbuckle. *Amos 6.0 User's Guide*. Amos Development Corporation, Springhouse, PA, 2005.

[6] D. Barnes, J. Carpenter, and D. Bailey. Partnerships with services users in interprofessional education for community mental health: a case study. *Journal of Interprofessional Care*, 14:191–202, 2000.

[7] Vic Barnett. *Sample Survey: Principles and Methods*. Arnold, London, 2002.

[8] H. Barr, M. Hamick, I. Koppel, and S. Reeves. Evaluating interprofessional education: two systematic reviews for health and social care. *British Educational Research Journal*, 25:533–543, 1999.

[9] P. R. Bernthal. Evaluation that goes the distance. *Training & Development*, pages 41–45, 1995.

[10] J. Bland and D. Altman. Multiple significance tests: the Bonferroni method. *British Medical Journal*, 310:170, 1995.

[11] K. A. Bollen. Indicator : Methodology. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social and Behavioral Sciences.*, pages 7282–7287. Oxford, 2001.

[12] Kenneth A. Bollen. *Structural Equations with Latent Variables.* Wiley, New York, 1989.

[13] Peter Bramley. *Evaluating Training Effectiveness, Translating Theory Into Practice.* McGraw-Hill, Maidenhead, 1991.

[14] R. Brinkerhoff. *Achieving results from training.* Jossey-Bass, San Francisco, 1987.

[15] D. Broomfield and J. Bligh. An evaluation of the 'short form' Course Experience Questionnaire with medical students. *Medical Education*, 32:367–369, 1998.

[16] D. Bushnell. Input, process, output: a model for evaluating training. *Training & Development*, 44:41–43, 1990.

[17] M. Byrne and B. Flood. Assessing the teaching quality of accounting programmes: an evaluation of the Course Experience Questionnaire. *Assessment and Evaluation in Higher Education*, 28:135–145, 2003.

[18] Alistair Campbell, Judith Walker, and Gerry Farrell. Confirmatory factor analysis of the GHQ-12: can I see that again? *Australian and New Zealand Journal of Psychiatry*, 37:474–483, 2003.

[19] John Carpenter. *Evaluating Outcomes in Social Work Education: Evaluation and Evidence, Discussion Paper 1.* Scottish Institute for Excellence in Social Work Education and the Social Care Institute for Excellence, January 2005.

[20] John Carpenter, Di Barnes, and Claire Dickinson. Making a modern mental health careforce: Evaluation of the Birmingham University Interprofessional Training Programme in Community Mental Health 1998-2002. Technical report, University of Durham, March 2003.

[21] Twinhead International Corp. *The Technology Beyond The Future.* `www.twinhead.com.tw`, 2005.

[22] J Cumming. *Clinical Decision Support.* PhD thesis, Department of Mathematical Sciences, Durham University, 2006.

[23] J. A. Cumming and D. A. Wooff. Dimension reduction via principal variables. *Computational Statistics & Data Analysis*, 52:550–565, 2007.

[24] Valbjorg Espeland and Oddny Indrehus. Evaluation of students' satisfaction with nursing education in Norway. *Journal of Advanced Nursing*, 42(3):226–236, 2003.

[25] Julian J. Faraway. *Practical Regression and Anova using R.* `www.stat.lsa.umich.edu/~faraway/book`, July 2002.

[26] Julian J. Faraway. *Linear Models with R.* Chapman & Hall/CRC, Boca Raton, USA, 2005.

[27] Peter M Fayers and David Machin. *Quality of Life: Assessment, Analysis and Interpretation.* Wiley, West Sussex, England, 2001.

[28] J. Fitz-Enz. Yes, you can weigh trainers' value. *Training*, 31:54–58, 1994.

[29] B Flury. *Common Principal Components and Related Multivariate Models.* Wiley, New York, 1988.

[30] J.D. Folley. The learning process. In *Training and Development Handbook.* McGraw-Hill, USA, 1967.

[31] Center for Social Epidemiology. *Job Stress Network.* `www.workhealth.org`, 2007.

[32] John Fox. Structural equation modeling with the sem package in R. *Structural Equation Modeling*, 13(3):465–486, 2006.

[33] David Garson. *Structural Equation Modeling.* `www2.chass.ncsu.edu/garson/pa765/structur.htm`, 2005.

[34] David Garson. *Structural Equation Modeling Example Using WinAMOS*. www2.chass.ncsu.edu/garson/pa765/semAMOS1.htm, 2005.

[35] David Goldberg. *Manual of the GHQ*. NFER Publishing Company, Windsor, 1978.

[36] Michael Goldstein and David Wooff. *Bayes Linear Statistics; Theory and Methods*. Wiley, Chichester, 2007.

[37] Franklin A. Graybill. *Matrices with applications in statistics, second edition*. Wadsworth, Belmont, California, 1983.

[38] A Hamblin. *Evaluation and control of training*. McGraw-Hill, London, 1974.

[39] Elwood F. Holton III. The flawed four-level evaluation model. *Human Resource Development Quarterly*, 7:5–21, 1996.

[40] Korporat INTAN. *Annual Report 2006*. The National Institute of Public Administration, Kuala Lumpur, 2007.

[41] J.J.Hox and T.M.Bechger. An introduction to structural equation modeling. *Family Science Review*, (11):354–373, 1998.

[42] Janice M Johnston, Gabriel M Leung, Richard Fielding, Keith Y K Tin, and Lai-Ming Ho. The development and validation of a knowledge, attitude and behaviour questionnaire to assess undergraduate evidence-based practice teaching and learning. *Medical Education*, 37:992–1000, 2003.

[43] K. G. Joreskog. A general method for estimating a linear structural equation system. In A. S. Goldberger and O. D. Duncan, editors, *Structural Equation Models in the Social Sciences*. Academic Press, New York, 1973.

[44] K. G. Joreskog. Structural equation models in the social sciences: specification, estimation and testing. In P. R. Krishnaiah, editor, *Applications of Statistics*. North-Holland, Amsterdam, 1977.

[45] Joseph F. Hair Jr., Rolph E. Anderson, Ronald L. Tatham, and William C. Black. *Multivariate Data Analysis (Fifth Edition)*. Prentice Hall, New Jersey, 1988.

[46] Winfred Arthur Jr., Winston Bennett Jr., Pamela S. Edens, and Suzanne T. Bell. Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology*, 88(2):234–245, 2003.

[47] R. S. Kaplan and D. P. Norton. Using the balanced scorecard as a strategic management system. *Harvard Business Review*, January/February:75–85, 1996.

[48] R. Kaufman, J. Keller, and R. Watkins. What works and what doesn't: Evaluation beyond Kirkpatrick. *Performance and Instruction*, pages 8–12, 1995.

[49] Paul Kearns and Tony Miller. *Measuring the impact of training and development on the bottom line*. Pearson Education, London, 1997.

[50] J. W. Keesling. *Maximum Likelihood Approaches to Causal Analysis*. PhD thesis, Department of Education, University of Chicago, 1972.

[51] E M Keramidas, S J Devlin, and R Gnanadesikan. A graphical procedure for comparing the principal components of several covariance matrices. *Communications in Statistics - Simulation and Computation*, 16:161–191, 1987.

[52] D. L. Kirkpatrick. Evaluation of training. In *Training and Development Handbook*. McGraw-Hill, USA, 1967.

[53] Jim Kirkpatrick. Transferring learning to behavior. *Training and Development*, 59(Issue 4):15, April 2005.

[54] Rex B. Kline. *Principles and Practice of Structural Equation Modeling*. Guilford, New York, 2005.

[55] K. Kraiger, J. K. Ford, and E. Salas. Application of cognitive, skill-based and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, 78:311–328, 1993.

[56] C. Kreber. The relationship between students' course perception and their approaches to studying in undergraduate science courses: a Canadian experience. *Higher Education Research and Development*, 22:57–75, 2003.

[57] W J Krzanowski and F H C Marriott. *Multivariate Analysis Part 1 : Distributions, Ordination and Inference.* Edward Arnold, London, 1994.

[58] Mark P. Leach and Annie H. Liu. Investigating interrelationships among sales training evaluation methods. *Journal of Personal Selling and Sales Management*, 23(4):327–339, fall 2003.

[59] Katholieke Universiteit Leuven. *ANCOVA*. `www.agr.kuleuven.ac.be/vakken/statisticsbyR/ANOVAbyRr/ANCOVAinR.htm`, July 2005.

[60] Richard Lowry. *One-Way Analysis of Covariance for Independent Samples Part 2.* `faculty.vassar.edu/lowry/ch17pt2.html`, 2005.

[61] RadhaKanta Mahapatra and Vincent S. Lai. Evaluating end-user training programs. *Communications of the ACM*, 48(1):67–70, January 2005.

[62] Economic Planning Unit Prime Ministers Department Malaysia. *Ninth Malaysia Plan 2006 - 2010: Efficient Public Service Delivery System.* `www.epu.jpm.my/rm9/english/Chapter26.pdf`, 2006.

[63] Richard McBain. Training effectiveness and evaluation. *Manager Update*, 15(3):23–34, Spring 2004.

[64] J. Moad. Industry outlook. *Datamation*, January:16–24, 1995.

[65] M. Molenda, J. A. Pershing, and C. M. Reigeluth. Designing instructional systems. In R. L. Craig, editor, *The ASTD training and development handbook 4th. ed.* McGraw-Hill, New York, 1996.

[66] A. Newby. *Training evaluation handbook.* Gower, Aldershot, 1992.

[67] Jum C. Nunnally and Ira H. Bernstein. *Psychometric Theory.* McGraw-Hill, New York, 1994.

[68] National Archive of Malaysia. *Pembukaan Pusat Latihan Pegawai-pegawai Kerajaan.* hids.arkib.gov.my/doc/jilidiii/september/19_09_1963_1980.htm, 2006.

[69] National Institute of Public Administration. *INTAN in brief.* www.intanbk.intan.my/cda/m_about/abt_brief.php, 2006.

[70] The National Institute of Public Administration. *INTAN in Brief.* The National Institute of Public Administration Malaysia, Kuala Lumpur, 2002.

[71] J. J. Phillips, editor. *In action series: measuring return on investment volume 1.* American Society for Training and Development, Alexandria, 1994.

[72] J. J. Phillips. Corporate training: does it pay off? *William & Mary Business Review*, pages 6–10, Summer 1995.

[73] H. Preskill and R. Torres. *Evaluative inquiry for learning in organizations.* Sage, Thousand Oaks, 1999.

[74] M. Pulley. Navigating the evaluation rapids. *Training & Development*, 48:19–24, 1994.

[75] Kia Fatt Quek, Wah Yun Low, Azad Hassan Razack, and Chit Sin Loh. Reliability and validiy of the general health questionnaire (ghq-12) among urological patients: A Malaysian study. *Psychiatry and Clinical Neurosciences*, 55:509–513, 2001.

[76] Leslie Rae. *How to measure training effectiveness; Second edition.* Gower, Hants, 1991.

[77] Paul Ramsden. A performance indicator of teaching quality in higher education: the Course Experience Questionnaire. *Studies in Higher Education*, 16(2):129–150, 1991.

[78] Margaret Anne Reid, Harry Barrington, and John Kenney. *Training Interventions, Managing Employee Development.* Institute of Personnel Management, Wimbledon, 1992.

[79] John T. E. Richardson. Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education*, 30:387–415, Aug. 2005.

[80] Dana Gaines Robinson and James C. Robinson. *Training for Impact: how to link training to business needs and measure the results.* Jossey-Bass Publishers, Oxford, 1991.

[81] Stephen Senn. *Statistical issues in drug development.* Wiley, Chichester, 1997.

[82] Takashi Shimizu, Tetsuya Mizoue, Hiroyuki Takahashi, Ayako Shazuki, Shinya Kubota, Norio Mishima, and Shoji Nagata. Relationships among self-management skills, communication with superiors, and mental health of employees in a Japanese worksite. *Industrial Health*, 41:335–337, 2003.

[83] C. Sleezer, D. Cipicchio, and D. Pitonvak. Customizing and implementing training evaluation. *Performance Improvement Quarterly*, 5:55–75, 1992.

[84] P. Sprent and N.C.Smeeton. *Applied Nonparametric Statistical Methods (Third Edition).* Chapman and Hall/CRC, London, 2001.

[85] James Stevens. *Applied Multivariate Statistics for the Social Sciences: Third Edition.* Lawrence Erlbaum Associates, New Jersey, 1996.

[86] D. Stufflebeam, W. Foley, et al. *L'evaluation en education et la prise de decision.* Edition NHP, Ottawa, 1980.

[87] P. Tamkin, J. Yarnall, and M. Kerrin. *Kirkpatrick and Beyond: a review of models of training evaluation.* The Institute for Employment Studies, Brighton, 2002.

[88] The R Development Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria, 2005.

[89] Jerry Trusty, Kok-Mun Ng, and Richard E. Watts. Model of effects of adult attachment on emotional emphaty of counseling students. *Journal of Counseling & Development*, 83:66–77, Winter 2005.

[90] Jodie B. Ullman. Structural equation modeling. In Barbara G. Tabachnick and Linda S. Fidell, editors, *Using Multivariate Analysis: Fourth edition.* Allyn and Bacon, USA, 2001.

[91] P. Warr, M. Bird, and N. Rackam. *Evaluation of management training.* Gower Press, London, 1970.

[92] B. L. Welch. On the comparison of several mean values: An alterntive approach. *Biometrika*, 38:330–336, 1951.

[93] U. Werneke, D.P.Goldberg, I. Yalcin, and B.T. Ustun. The stability of the factor structure of the General Health Questionnaire. *Psychological Medicine*, 30:823–829, 2000.

[94] M. Whitelaw. *The evaluation of Management Training: a Review.* Institute of Personnel Management, London, 1972.

[95] D. E. Wiley. The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger and O. D. Duncan, editors, *Structural Equation Models in the Social Sciences.* Academic Press, New York, 1973.

[96] Keithia L. Wilson, Alf Lizzio, and Paul Ramsden. The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education*, 22(1):33–53, 1997.

[97] WolframMathWorld. *Bonferroni Correction.* `mathworld.wolfram.com/BonferroniCorrection.html`, 2007.

[98] Karl L. Wuensch. *Review of articles of use exploratory factor analysis.* `core.ecu.edu/psyc/wuenschk/StatHelp/EFA.htm`, 2005.

[99] C.V. Youmans. Testing for training and development. In *Training and Development Handbook.* McGraw-Hill, USA, 1967.

Appendix A

Pretest Learning Questionnaire

EVALUATION OF TRAINING EFFECTIVENESS

Tuan/Puan,

You have been selected as a participant in exploratory research on the evaluation of training effectiveness. This research will contribute towards a more structured and effective way of measuring the performance of training programmes.

There are two sets of questionnaires for you to answer. The pre-test set consists of the General Health questionnaire and the Learning questionnaire. The post-test set, which will be distributed at the end of your training programme, consists of the Course Experience Questionnaire and the Learning questionnaire.

Throughout the research your individual score will remain confidential. Analysis will only be done on the aggregate scores of the group, which forms the main interest of the research. Your kind cooperation in completing this set and the set after the course is very much appreciated.

Thank you.

Sincerely,

Anesee Ibrahim

Your course code: □□□□□□

Your respondent code: □□□□□□

Personal information : Please tick (/) or cross (X) your answers.

Gender:

□ Male
□ Female

Ethnic origin:

□ Malay
□ Chinese
□ Indian
□ Others

Age:

□ Below 26 years
□ 26 – 30
□ 31 – 35
□ 36 – 40
□ 41 – 45
□ Above 45 years

Type of service:

□ Federal
□ State
□ Local Government
□ Statutory body
□ Others

Group of post:

□ P & P
□ Supporting 1
□ Supporting 2

Years in service:

□ 5 years or less.
□ 6 to 10 years.
□ 11 to 15 years.
□ More than 15 years.

GENERAL HEALTH QUESTIONNAIRE

We would like to know if you have had any medical complaints and how your health has been in general over the past few weeks. Please answer ALL of the following questions simply by ticking (/) or crossing (X) the answer you think most nearly applies to you. Remember that we want to know about present and recent complaints, not those you had in the past. Your answers to the questions will remain confidential. We are not at all interested in identifying you personally, rather we are interested in the overall levels of health in the group.

HAVE YOU RECENTLY:

1. Been able to concentrate on whatever you are doing?	□ Better than usual	□ Same as usual	□ Less than usual	□ Much less than usual
2. Lost much sleep over worry?	□ Not at all	□ No more than usual	□ Rather more than usual	□ Much more
3. Felt that you are playing a useful part in things?	□ More so than usual	□ Same as usual	□ Less useful than usual	□ Much less than usual
4. Felt capable of making decisions about things?	□ More so than usual	□ Same as usual	□ Less capable than usual	□ Much less capable
5. Felt constantly under strain?	□ Not at all	□ No more than usual	□ Rather more than usual	□ Much more than usual
6. Felt you couldn't overcome your difficulties?	□ Not at all	□ No more than usual	□ Rather more than usual	□ Much more than usual
7. Been able to enjoy your normal day-to-day activities?	□ More so than usual	□ Same as usual	□ Less so than usual	□ Much less than usual
8. Been able to face up to your problems?	□ More so than usual	□ Same as usual	□ Less able than usual	□ Much less able
9. Been feeling unhappy and depressed?	□ Not at all	□ No more than usual	□ Rather more than usual	□ Much more than usual
10. Been losing confidence in yourself?	□ Not at all	□ No more than usual	□ Rather more than usual	□ Much more than usual
11. Been thinking of yourself as a worthless person?	□ Not at all	□ No more than usual	□ Rather more than usual	□ Much more than usual
12. Been feeling reasonably happy, all things considered?	□ More so than usual	□ About same as usual	□ Less so than usual	□ Much less than usual

LEARNING QUESTIONNAIRE

This questionnaire has three (3) parts. Each part refers to eight (8) subject areas, plus two aspects related to the area targetted by the training programme. On this page, Part A measures your perception towards your own level of knowledge. Part B and Part C are on the following pages.

Part A : Level of Knowledge

This part of the questionnaire measures your perception towards your own **level of knowledge** in each of the following subject areas. Please think of your level of knowledge in each of the subject areas with regard to the general average level of knowledge among your **colleagues of the same rank**. Please indicate your choice by ticking (/) or crossing (X) in a box for each subject based on the Low - High scale below.

	Low ←						→ High
1. Economic Management	□	□	□	□	□	□	□
2. Financial Management	□	□	□	□	□	□	□
3. Information Technology and Communication	□	□	□	□	□	□	□
4. Human Resource and Organisation	□	□	□	□	□	□	□
5. Social and Infrastructure Planning and Administration	□	□	□	□	□	□	□
6. Land, Territorial, Regional and Local Government Administration	□	□	□	□	□	□	□
7. International Relations and Foreign Affairs	□	□	□	□	□	□	□
8. Defence and National Security	□	□	□	□	□	□	□

*For questions 9 and 10, please evaluate your **level of knowledge and skill** in the area targeted by the training programme that you are about to attend.*

9. Knowledge	□	□	□	□	□	□	□
10. Skill	□	□	□	□	□	□	□

Thank you. Please move on to the following page to Part B of the Learning questionnaire.

LEARNING QUESTIONNAIRE

Part B : Application and use.

When you work there is specific knowledge that you use in order to do your job effectively. For each of the subject areas below, please indicate whether the subject is highly used or not used at all, or any other positions in between, by ticking (/) or crossing (X) in a box that you think appropriate.

This refers to the knowledge that you use when you carry out your **main responsibility at your workplace**.

	Not used at all ←					→ Highly used
1. Economic Management	□	□	□	□	□	□	□
2. Financial Management	□	□	□	□	□	□	□
3. Information Technology and Communication	□	□	□	□	□	□	□
4. Human Resource and Organisation	□	□	□	□	□	□	□
5. Social and Infrastructure Planning and Administration	□	□	□	□	□	□	□
6. Land, Territorial, Regional and Local Government Administration	□	□	□	□	□	□	□
7. International Relations and Foreign Affairs	□	□	□	□	□	□	□
8. Defence and National Security	□	□	□	□	□	□	□

*For questions 9 and 10, please indicate the **level of usage of the knowledge and skill** in the area targeted by the training programme that you are about to attend.*

9. Knowledge	□	□	□	□	□	□	□
10. Skill	□	□	□	□	□	□	□

Thank you. Please move on to the following page to Part C of the Learning questionnaire.

LEARNING QUESTIONNAIRE

Part C: Importance of learning

How important is it for you personally **to continue learning and enhancing your knowledge** in each of the subject areas? Again please indicate the level of importance by choosing a box based on the importance scale below.

Not important at all ← → Highly important

1. Economic Management	□	□	□	□	□	□	□
2. Financial Management	□	□	□	□	□	□	□
3. Information Technology and Communication	□	□	□	□	□	□	□
4. Human Resource and Organisation	□	□	□	□	□	□	□
5. Social and Infrastructure Planning and Administration	□	□	□	□	□	□	□
6. Land, Territorial, Regional and Local Government Administration	□	□	□	□	□	□	□
7. International Relations and Foreign Affairs	□	□	□	□	□	□	□
8. Defence and National Security	□	□	□	□	□	□	□

For questions 9 and 10, please indicate the importance of learning the ***knowledge and skill*** *in the area targeted by the training programme that you are about to attend.*

9. Knowledge	□	□	□	□	□	□	□
10. Skill	□	□	□	□	□	□	□

Your have completed all three parts of the Learning questionnaire. Thank you very much for your kind cooperation.

Appendix B

Posttest Learning Questionnaire

COURSE EXPERIENCE QUESTIONNAIRE

Your course code: □□□□□□

Your respondent code: □□□□□□

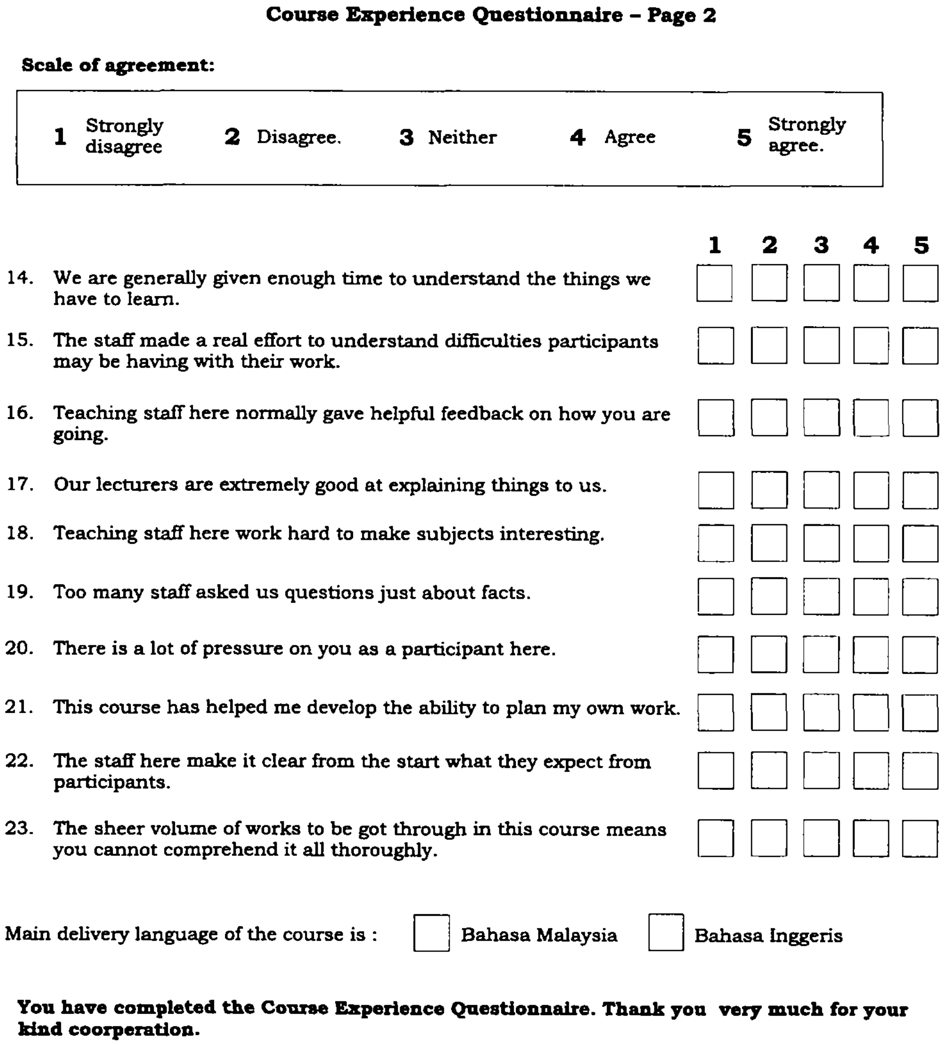
The following 23 statements refer to the course or training programme that you have just attended. For each statement, please respond by ticking (/) or crossing (X) in the box that correctly represents your agreement, based on the five point scale below. You have to respond to all statements.

Your response to the statements will remain confidential. We are not at all interested in identifying you personally, rather we are interested in the overall levels of experience in the programme. Your cooperation is very much appreciated.

Scale of agreement:

1 Strongly disagree	**2** Disagree.	**3** Neither	**4** Agree	**5** Strongly agree.

Statements:	**1**	**2**	**3**	**4**	**5**
1. It is always easy to know the standard of work expected.	□	□	□	□	□
2. The course has helped me to develop my problem-solving skills.	□	□	□	□	□
3. The teaching staff of this course motivates participants to do their best work.	□	□	□	□	□
4. The workload is too heavy.	□	□	□	□	□
5. The course has sharpened my analytical skills.	□	□	□	□	□
6. You usually have a clear idea of where you are going and what is expected of you.	□	□	□	□	□
7. Staff here put a lot of time in commenting on participants' work.	□	□	□	□	□
8. To do well on this course all you really needed was a good memory.	□	□	□	□	□
9. This course has helped develop my ability to work as a team member.	□	□	□	□	□
10. As a result of doing this course, I feel more confident about tackling unfamiliar problems.	□	□	□	□	□
11. This course has improved my written communication skills.	□	□	□	□	□
12. Staff seemed more interested in testing what you have memorized than what you have understood.	□	□	□	□	□
13. It is often hard to discover what is expected of you in this course.	□	□	□	□	□



Course Experience Questionnaire - Page 2

Scale of agreement:

1 Strongly disagree	**2** Disagree.	**3** Neither	**4** Agree	**5** Strongly agree.

		1	2	3	4	5
14.	We are generally given enough time to understand the things we have to learn.	□	□	□	□	□
15.	The staff made a real effort to understand difficulties participants may be having with their work.	□	□	□	□	□
16.	Teaching staff here normally gave helpful feedback on how you are going.	□	□	□	□	□
17.	Our lecturers are extremely good at explaining things to us.	□	□	□	□	□
18.	Teaching staff here work hard to make subjects interesting.	□	□	□	□	□
19.	Too many staff asked us questions just about facts.	□	□	□	□	□
20.	There is a lot of pressure on you as a participant here.	□	□	□	□	□
21.	This course has helped me develop the ability to plan my own work.	□	□	□	□	□
22.	The staff here make it clear from the start what they expect from participants.	□	□	□	□	□
23.	The sheer volume of works to be got through in this course means you cannot comprehend it all thoroughly.	□	□	□	□	□

Main delivery language of the course is : □ Bahasa Malaysia □ Bahasa Inggeris

You have completed the Course Experience Questionnaire. Thank you very much for your kind coorperation.

LEARNING QUESTIONNAIRE

This questionnaire has three (3) parts. Each part refers to eight (8) subject areas, plus two aspects related to the area targetted by the training programme. On this page, Part A measures your perception towards your own level of knowledge. Part B and Part C are on the following pages.

Part A : Level of Knowledge

This part of the questionnaire measures your perception towards your own **level of knowledge** in each of the following subject areas. Please think of your level of knowledge in each of the subject areas with regard to the general average level of knowledge among your **colleagues of the same rank**. Please indicate your choice by ticking (/) or crossing (X) in a box for each subject based on the Low - High scale below.

Low ← → High

1. Economic Management □ □ □ □ □ □ □

2. Financial Management □ □ □ □ □ □ □

3. Information Technology and Communication □ □ □ □ □ □ □

4. Human Resource and Organisation □ □ □ □ □ □ □

5. Social and Infrastructure Planning and Administration □ □ □ □ □ □ □

6. Land, Territorial, Regional and Local Government Administration □ □ □ □ □ □ □

7. International Relations and Foreign Affairs □ □ □ □ □ □ □

8. Defence and National Security □ □ □ □ □ □ □

For questions 9 and 10, please evaluate your ***level of knowledge and skill*** *in the area targeted by the training programme that you have just attended.*

9. Knowledge □ □ □ □ □ □ □

10. Skill □ □ □ □ □ □ □

Thank you. Please move on to the following page to Part B of the Learning questionnaire.

LEARNING QUESTIONNAIRE

Part B : Application and use.

When you work there is specific knowledge that you use in order to do your job effectively. For each of the subject areas below, please indicate whether the subject is highly used or not used at all, or any other positions in between, by ticking (/) or crossing (X) in a box that you think appropriate.

This refers to the knowledge that you use when you carry out your **main responsibility at your workplace**.

Not used at all ←——— ———→ Highly used

1. Economic Management □ □ □ □ □ □ □

2. Financial Management □ □ □ □ □ □ □

3. Information Technology and Communication □ □ □ □ □ □ □

4. Human Resource and Organisation □ □ □ □ □ □ □

5. Social and Infrastructure Planning and Administration □ □ □ □ □ □ □

6. Land, Territorial, Regional and Local Government Administration □ □ □ □ □ □ □

7. International Relations and Foreign Affairs □ □ □ □ □ □ □

8. Defence and National Security □ □ □ □ □ □ □

*For questions 9 and 10, please indicate the **level of usage of the knowledge and skill** in the area targeted by the training programme that you have just attended..*

9. Knowledge □ □ □ □ □ □ □

10. Skill □ □ □ □ □ □ □

Thank you. Please move on to the following page to Part C of the Learning questionnaire.

LEARNING QUESTIONNAIRE

Part C: Importance of learning

How important is it for you personally **to continue learning and enhancing your knowledge** in each of the subject areas? Again please indicate the level of importance by choosing a box based on the importance scale below.

Not important at all ← → Highly important

1. Economic Management □ □ □ □ □ □ □

2. Financial Management □ □ □ □ □ □ □

3. Information Technology and Communication □ □ □ □ □ □ □

4. Human Resource and Organisation □ □ □ □ □ □ □

5. Social and Infrastructure Planning and Administration □ □ □ □ □ □ □

6. Land, Territorial, Regional and Local Government Administration □ □ □ □ □ □ □

7. International Relations and Foreign Affairs □ □ □ □ □ □ □

8. Defence and National Security □ □ □ □ □ □ □

*For questions 9 and 10, please indicate the importance of learning the **knowledge and skill** in the area targeted by the training programme that you have just attended.*

9. Knowledge □ □ □ □ □ □ □

10. Skill □ □ □ □ □ □ □

Your have completed all three parts of the Learning questionnaire. Thank you very much for your kind cooperation.