DETERMINING RELATIVE ACADEMIC ACHIEVEMENT FOR FAIR ADMISSION TO HIGHER EDUCATION

Report to a Joint Committee of the Australian National University,
the Canberra College of Advanced Education, and the
Australian Capital Territory Schools Authority, appointed to
Supervise Research into Tertiary Entrance Score Calculations

D. J. DALEY

Statistics Department

School of Mathematical Sciences

Australian National University

January 1989

This copy of the Report has been reset from T3 into TEX. The only changes made have been the correction of typographical errors, and Table 12.4 which was based on an incorrect table from an earlier manuscript. The pagination here is more compressed (cf. the Table of Contents).

Preface

Tertiary admission for immediate school leavers in Australia has been selective for a couple of decades. Selection has been based on "relative academic merit" as calculated by a variety of methods from either or both of public examination marks and school-based assessments. It is not easy to locate documents describing the computations used in the different states, and it is even harder to find self-contained and self-consistent accounts of the rationale used to justify the numerical procedures that are followed. Cooney's (1976) review is not of much assistance on matters of detail. What is readily acknowledged is that "statistically based procedures" are used, but as e.g. Masters and Beswick (1986, §2.32) observed,

> "A weakness of scaling procedures in use in every system in Australia is that they are not supervised by an explicit statistical model for score equating and aggregation."

It is almost as difficult to ascertain what different perceptions the various educational authorities have as to what is involved in determining "relative academic merit". To a fair extent, methods have evolved on an *ad hoc* basis, as the quotation above may indicate. Where all systems agree is that there are technical considerations involved, though in the case of the Australian Capital Territory the attitude to this aspect has been largely to keep the technical people distant from the decision making bodies, as if the former are unable to appreciate educational principles. Two specific statements from the more deliberative venue of reports are worth noting:

> "Simplicity may be beguiling . . . Statistical complexity may obfuscate the process for the public . . . The critical consideration, however, should not be simplicity or complexity but fairness." (*Making Admission to Higher Education Fairer*, §4.19)

> "It is sometimes assumed in the education field, if not others, that the adequacy of a system is measured by the ready comprehensibility of its technical details . . . The Working Party has chosen the approach that what should be readily understood are the principles on which the system is based . . . It has proved impossible to devise a system that both discharges some worthwhile principles and also has no technical aspects of greater complexity than can be explained to a general audience in a few words." (*Tertiary Entrance in Queensland: A Review*, Report of the Working Party on Tertiary Entrance, Board of Secondary School Studies, Brisbane, July 1987, p.123.)

My background in the language of mathematics and statistics is more extensive than that of some of my predecessors who have written reports on TE score construction for the ACT Schools Accrediting Agency. For this reason and because I have been asked for a statement of the theoretical basis for the Method-of-Moment scaling procedure, I address – and gladly – some of the technical issues behind the construction of Tertiary Entrance scores. From this stance, some previous reports have aspects which are deficient, so given my brief I have noted some of these shortcomings.

A general attitude that motivates our discussion is an attempt to understand and describe simple general mathematical structures such as underlie the Queensland TE score algorithm on which the ACT's algorithm is based. In the report, I use such a structure to describe idealized versions of data sets such as ACT students' course scores. I explore the consequences of assuming, first that the data have this structure, and then that they deviate from it in as extreme a manner as is consistent with the type of data sets we actually see. In pragmatic terms this exploration shows that the initial simplifying assumptions do not have significant consequences for the computations devised assuming that the model is "true".

Our discussion is starts from a set of basic assumptions and principles which guide our analyses and the development of our procedures. Chapter 1 is an introduction and extended summary containing no mathematics. It includes our recommendations. Any disputation of these conclusions should be based on both the initial premisses and the body of the report. While it may be daunting to be presented with "conclusions" before the evidence and/or argument has been given, it is one way of providing the lay reader with a lengthier account than the more tersely written non-technical summary. This is followed by the various Conclusions that are stated more formally through the report.

Subsequent chapters have much that is devoid of any algebra, if this is what is regarded as being "technical" language. We try to follow a logical development, and attempt to underline the significance of any conclusions that are not self-evident. Certain material that is either more technical or not readily available is included in various appendices, in the hope that they may be of at least the same use as appendices in other reports that I have used over the past few years.

At the outset it was clear that all data analyses should be executed on a personal computer. I looked for compatibility with IBM PC's as these are generally available and are well suited to the intensive numerical work entailed in some of the operations required in a study of a scaling procedure.

The ACT Schools Accrediting Agency supplied[1] me with 1986 data (course, ASAT, and TE score information) in August 1987. These fitted onto three standard PC diskettes. The data files were used in a still more compact format to facilitate rapid execution: even for the largest college (the *college* was the module for most computational purposes), it was possible to compress the data enough to run all the Fortran programmes written for non-standard analyses without exceeding the data default size limitation of 64Kb, enabling speedy numerical processing with minimal *input/output* delays. I have used IBM PC AT computers or clones of them with 640K RAM and 80287 coprocessor; less RAM may suffice, but I have not tested this aspect. I used the statistical package SYSTAT for several routine procedures.

I thank many people for assisting in the development of this report. Ms. Yvonne Pytelkow helped in transcribing the data supplied on tape via Mr. R. Edwards of the ACT Schools Accrediting Agency; he also gave other detailed information, and after three iterations the experimental routines developed for studying the scaling procedure yielded input parameters for him to use in existing computer programs of the ACT Schools Authority that lead to the annual *Year 12 Study*. Mr. Reg Allen of the Queensland BSSS has been a helpful and sympathetic colleague in the later stages, as also was Mr. George Morgan of ACER. Mr. J. C. Daley has given invaluable editorial assistance. The work was begun while visiting the University of North Carolina at Chapel Hill and concluded in Canberra.

Finally, I pay tribute here to the quiet tenacity of Mrs. Joan Robson, formerly of St. Clare's College. Much in this report has come about because of her intuition, data-based insights, and insistence from 1977 on that all is not well with the ACT TE score.

D. J. Daley, Canberra, January 1989

[1] Tapes of 1987 data have been provided on 15 December 1988, but with different content from the 1986 tapes. I was given a second set of tapes on 13 January 1989, but still missing some information. Several analyses summarized in this report should be replicated on 1987 data — and maybe 1988 as well — and noted in a sequel.

Contents*

Preface iii (iii)

Non-technical Summary viii (xi)

Conclusions xi (xv)

1. Introduction and Summary 1 (1)

The Project as Assigned

Some General Background

Principles and Assumptions for Producing Tertiary Entrance Scores in the A.C.T.

Basic Assumptions

Principles in Constructing TE Scores

Assumptions about ASAT Scores

A Logical Basis for Scaling

What is a Scaling Procedure?

First Approaches

Data Analyses

Second Approaches

APPENDIX TO CHAPTER 1:

The Nature of the Project 9 (13)

2. Representations of Course Scores and the Construction of Aggregates 11 (16)

What an Aggregate Score Represents

When is an Aggregate a "Fair" Aggregate?

What do Scores Represent? Another Reason for using Aggregates

Brief Review and Preview

MATHEMATICAL APPENDIX TO CHAPTER 2:

Determination of the Scale Parameters: So-called Iterative Scaling Procedures 18 (25)

3. Modelling Course Scores 20 (28)

True Score Modelling

One-factor Models and Parameter Estimation

Method-of-Moment Estimation

Two-factor Models

* The original pagination is given in parentheses.

4. Constructing and Using an External Reference Scale 25 (34)

Constructing an Optimum Reference Scale
Using Reference Scale Scores
True Score Modelling
One-factor Model
Interlude: Principal Component Representations
One-factor Model (cont.)
Weight of ASAT Scores in a One-factor Model
Educational Measurement Approach
The Outlier Problem
Several Reference Scales?

5. Properties of ASAT Scores 42 (57)

Measurement Error Variances
Psychometric Stability of ASAT Scores
Relative Structure of Quantitative, Verbal and Writing Task Scores

6. Discrepancies Between ASAT and Course Scores 48 (65)

Gender-linked Mode of Assessment Discrepancy
Use of Mode of Assessment Discrepancy Measures
Writing Task and Multiple Choice Verbal Scores

7. Bias Problems with Reference Scales 56 (75)

The Original Complaint
Some Average Achievement and Test Scores
Masters and Beswick's Evidence
Gender-linked Biases and Other Course Score Scaling

8. Imprecision in Calculating Tertiary Entrance Scores 65 (87)

Scope for Jackknife Studies
Interpretation

9. The A.C.T. Tertiary Entrance Score: Background and Prescriptive Structure 72 (95)

The Source Data
The Existing Scaling Procedure
Moderation Group Assumptions
Structure of TE Scores and Other Course Score Scaling Criteria
Some Possible Abuses
Prescriptive Actions and ASAT Scores

10. Some A.C.T. Course Score Analyses 78 (103)

Three-factor Structure
Two-factor Structure of Course Scores

11. An Optimally Constructed A.C.T. Tertiary Entrance Score 84 (110)

12. Some Comparisons of Three Scaling Procedures via TE Scores and other Parameters 87 (114)

Some Examples of Different TE Scores
Rankings from Different TE Scores
Moments of Different TE Score Distributions
Regression Study showing TEACT/TEMM Difference
Moderation Group Parameters

13. Miscellanea 107 (137)

Scaling Procedures and Algorithms
Why Simulation Studies?
The 1977–85 ACT Algorithm — Model 1
The 1977–85 ACT Algorithm — Model 2
Method-of-Moment Scaling
Multiple Aggregates

References 114 (147)

Appendix 1: *Reprint of* Modelling Examination Marks (II) 116 (150)

Appendix 2: *Reprint of* Different sex differences from different modes of assessment: common experiences in three countries 131 (178)

Non-technical Summary

Cautionary remark. This report is about scaling scores that can be school-based assessments of relative achievement, or examination marks, or a combination of both. To the extent that this topic has a strong technical aspect—hence this non-technical summary—neither this short summary nor the introduction is any substitute for the report proper when it comes to matters of argument.

"It is sometimes assumed in the educational field, if not others, that the adequacy of a system is measured by the ready comprehensibility of its technical details . . . [This] Working Party has chosen the approach that what should be readily understood are the principles on which the system is based . . . It has proved impossible to devise a system that both discharges some worthwhile principles and also has no technical aspects of greater complexity than can be explained to a general audience in a few words." (pp. 123–124 of *Tertiary Entrance in Queensland: A Review*, July, 1987)

The essence of this report is the provision of advice concerning the methodology and procedures that are or might be used to construct a Tertiary Entrance (TE) score in the Australian Capital Territory (ACT).

The report starts in Chapter 1 with a statement of Principles and Assumptions that supposedly underlie the current procedure. All except the last either reflect policy principles or are consistent with properties of data representing relative student achievement in the upper secondary school level curriculum, whether from the ACT or Queensland or New South Wales or any other Australian state or . . . , and whether school-based or public examination based or a combination of both.

Chapter 1 gives a general introduction and longer summary of the report, non-technical in the sense of no algebra. Chapter 2 gives a theoretical analysis of an idealized data set which is complemented in Chapter 3 by a description of a model that is more flexible and copes with real data sets. The combination of these two gives a framework within which the Principles listed at the outset can be translated into a practical algorithm or numerical procedure to produce a Tertiary Admissions Index like the present TE score, within any college. It leads to the conclusion, supported by empirical work in Chapters 7, 8 and 10, that the present method for producing TE scores is unnecessarily shoddy, reflecting its origins as a "quick-fix" solution to a problem. It is made shoddy by its excessive dependence on Australian Scholastic Aptitude Test (ASAT) scores. Examples of three "TE scores" in Chapter 12 show that this shoddiness has consequences for appreciable proportions of students admitted by tertiary institutions. The major source of this shoddiness is the discrepancy between ASAT scores as measures of "developed general ability" and the school-based scores used to give a measure of "relative general achievement".

The ACT and Queensland systems are distinguished from the rest of Australia by an absence of system-wide public examinations. As substitutes for establishing some system-wide measures of academic achievement, ASAT sub-scale and (more recently) Writing Task scores are used in a statistical scaling procedure that is not as consistent with the policy principles as an Optimal Other Course score Scaling Procedure that is described in Chapter 11 up to limits imposed by a Research Supervisory Committee. In Chapter 4 we show how to find an optimal mixture of these three sub-scale scores for use as a reference scale for all colleges.

Finally, the effect of producing Tertiary Admission Indices by such a better method is shown to have little effect on the more basic and transparent sex bias inequity that has plagued TE scores since their inception in 1977, despite various relatively ineffectual attempts to rectify the matter in 1983, 1984 and 1986. A statistical remedy is at present the only practicable one, yet successive Committees have shied away from it seemingly from fear of upsetting a public that may then clamour for the return of greater external assessment and accountability of the system.

Like Masters & Beswick (1986)[2], I recognize that the assumption that a one-factor model describes the data adequately is open to question. Such a model is purportedly the basis of the existing scaling procedure or *algorithm* which has been used with various modifications to produce ACT TE scores in the period 1977–87. Both theoretically and empirically, that algorithm is shoddy relative to an optimal procedure based on the same model but having superior properties with respect to both unbiasedness and precision (meaning here, faithfulness to the data within the limits attainable due to basic "measurement errors").

Unlike Masters & Beswick, I have asked what are the effects of assuming a more complex multidimensional model description for the data, and used tools of theory, "typical" data parameters, and empirical studies to answer that question. The interpretation of a Tertiary Entrance score as a (relative) measure of general academic achievement can be retained, albeit that it is now to be viewed as a measure that reflects, approximately uniformly for all students, a mixture of both their general achievement measure and the extent of departure of their achievements in the area(s) where these may be stronger and more specialized. Neither the theory nor the empirical studies can produce evidence of the introduction of biases through the statistically based procedure as suggested but not proved in any way by Masters and Beswick (their empirical analyses were fundamentally flawed by data-selection effects and taking no account of the actual structure of the data, in spite of having shown they were aware of part of that structure when they wrote a report in 1985). In other words, using the optimal algorithm for constructing Tertiary Entrance scores on the basis of the one-factor model but applying it to a multidimensional model with "typical" data parameter values, makes little difference to the fairness of the general achievement measure being produced.

What emerges from combining the empirical and theoretical studies is a reinforcement of a recommendation of 1985 to the ACT Schools Accrediting Agency or Secretariat that (1) the existing algorithm for calculating two parameters for each course (= *moderation group*) so as to "place all course scores on a common scale" should be replaced by the superior method-of-moment procedure. This requires also (2) the use of suitable statistical techniques to enforce compliance by ASAT scores with the assumptions made about their joint properties with course scores, as without such compliance the claims currently made for the integrity of TE scores constitute gross exaggerations.

The effect of these recommendations will be the production of TE scores which have considerably smaller biases and increased precision, and are also more faithful to the proclaimed policy dictum that "ACT TE scores reflect school-based assessments". If there is belief in this policy dictum, then the existing scaling procedure should be scrapped immediately in favour of one based on the Method-of-Moment procedure, coupled with statistical measures to control the excessive deviations of ASAT scores: it has been amply demonstrated over the past decade that these excesses are beyond control by the prescriptive methods which have been tried *ad nauseam.*

Some of the conclusions which the analyses of the data have thrown up may be as unpalatable as the existence of a gender-linked sex bias between ASAT and school-based assessments has proved to be (and, as an aside, note that the bias persisted in 1986 and 1987, albeit at levels lower than what they would have been without the introduction of a Writing Task). Whatever, the data can only be allowed to speak for themselves, and I have not seen my task as involving the deliberate inclusion or exclusion of information whose content is unaltered by wishful thinking. In any case, the conclusions reached in this report via analyses on full data sets, simply confirm much of what can be deduced from the summary data published each year in the *Year 12 Study.*

"*Simplicity may be beguiling . . . Statistical complexity may obfuscate the process for the public . . . The critical consideration, however, should not be simplicity or complexity but fairness.*" (*MATHEF*, §4.19).

[2] G. N. Masters & D. G. Beswick (1986). *The Construction of Tertiary Entrance Scores: Principles and Issues.* Centre for the Study of Higher Education, University of Melbourne.

DETERMINING RELATIVE ACADEMIC ACHIEVEMENT FOR FAIR ADMISSION TO HIGHER EDUCATION

"CONCLUSION" STATEMENTS

1.1. If a composite measure of relative general academic achievement is to be constructed on the basis of several particular measures of achievement like exam. marks, then

(A) it is statistically acceptable that it be an aggregate of scaled marks; and

(B) the scaling should be determined internally by the set of marks in conjunction with the method-of-moment estimation procedure;

(C) an aggregate of marks determined via (B) is also educationally acceptable.

2.1. A ranking determined by aggregates from data sets such as $\mathcal{X}$ is basically a ranking determined by the first principal components $\{U_{i1}\}$ in the principal component representation.

REMARK 2.1. Conclusion 2.1 does not depend on any of the statistical modelling assumptions to be discussed in Chapter 3.

2.2. When all students follow a common curriculum, the construction of a "best m subset" aggregate from course scores $\mathcal{X} \equiv \{X_{ij}\}$ most faithfully represents the principle P 1 that students' curriculum choices should not affect their TE scores when student $i's$ scores $X_i = (X_{i1} \;\cdots\; X_{in})'$ are rescaled to scores $Y_i = (Y_{i1} \;\cdots\; Y_{in})'$ via relations

$$Y_{ij} = \beta_j X_{ij}$$

for constants β_j such that the matrix $B \sum_X B$ with $B = \mathrm{diag}(\{\beta_j\})$ and $\sum_X$ the covariance matrix of $\mathcal{X}$ has $e/\sqrt{n} \equiv (1/\sqrt{n} \cdots 1/\sqrt{n})'$ for the eigenvector associated with its largest eigenvalue.

2.3. The scaling constants β_j of Conclusion 2.2 are determined uniquely apart from a multiplicative constant, and there is a convergent iterative procedure to determine them.

2.4. Constructing a general achievement index from independence assumptions and a quasi-Rasch model that uses course scores as parameters leads to the same starting point as the beginning of Chapter 2.

3.1. The Method-of-Moment estimation procedure for finding scale parameters yields the same estimators as under a principal component analysis when the data set is a balanced set as in Chapter 2.

3.2. For practical purposes, fitting a data set having a two-factor structure as at (3.14) by the Method-of-Moment estimation procedure valid for scaling a one-factor model results in negligible differences in the aggregate scores produced from the original and scaled data sets.

4.1. In terms of providing a better reference scale, the introduction of the Writing Task in 1986 was a success.

4.2. Each year, the optimal predictive combination of system-wide reference scores should be constructed to produce a single system-wide reference measure of general ability in the "dimension" of school-based general achievement measures.

4.3. The coherence between course scores and even optimally constructed “Total ASAT score” to be used as a reference scale, varies significantly between colleges.

REMARK 4.1. The ACT system already uses Other Course Score scaling criteria.

4.4. Reference scale scores and course scores should be treated as a single data set for scaling purposes using the Method-of-Moment estimation procedure to find the scale parameters $(\beta\ \beta_A)$ with a weight factor w_A attached to the reference scale scores given by (4.29), and the normalization constant for $(\beta\ \beta_A)$ determined by setting $\beta_A = 1$.

4.5. The multivariate data set approach of Conclusion 4.3 coincides with the two-stage approach of constructing school-based estimates of relative general achievement within a college, and subsequently combining these optimally with an external set of reference scores to produce system-wide estimates of relative general achievement.

4.6. The one-step procedure of the Method-of-Moment estimation procedure can be dissected so as to furnish colleges, *if they so desire*, with *approximate* “within-college” estimates of their students’ relative general achievements or “within-college” TE scores, at any time that a school-based set of quasi-course scores is known. In particular, a purely school-based set of general achievement scores can be furnished, but such scores would have no between-school comparability. Such estimates would be subject to minor adjustment at a final stage when system-wide reference scores are determined and made known.

4.7. The present *statistical* use of ASAT scores lacks attention to critical detail. Scrutiny shows that the data deviate excessively from the assumed close relationships. Without these relationships, further checks and adjustments are necessary to justify the statistical use of ASAT scores.

4.8. In terms of both face validity and modelling considerations, Other Course Score scaling is both consistent with a model aimed at producing a single aggregate and with yielding an aggregate free(r) of bias from selection effects.

5.1. The units of the integer-valued sub-scale scores Q , V and W corresponded in 1986 to 12%, 17% and 30% of their respective standard deviations. The measurement error standard deviations of these scores are about 4 units for Q and V, and 2 to 3 units for W.

5.2. For the purpose for which ASAT scores are used in the ACT, neither they nor the Quantitative and Verbal sub-scale scores are psychometrically stable with respect to gender differences in different years.

6.1. Students’ relative *abilities* as assessed by multiple choice methods in the ASAT test and reported on Quantitative and Verbal sub-scales, are positively correlated with but differ systematically from school-based determinations of their relative *achievementsin* the related areas of Mathematics and English respectively, this difference being a characteristic of the two modes of assessment used in the two pairs of scores. Irrespective of the Quantitative or Verbal “dimension” concerned, females tend to perform relatively better on the school-based measures and males on the ASAT test.

6.2. On the basis of the TE score construction principles P 2–4, Conclusion 6.1 implies that ASAT scores are biased for use as reference scores in the ACT. If similar principles hold in Queensland, the same conclusion holds there also.

6.3. Analysis based on a linear representation gives no evidence of association between the gender-linked discrepancy between course and ASAT scores and the known gender-linked difference in verbal and quantitative skills.

6.4. Test results for both 1986 and 1987 revealed systematic differences in assessment by essay-writing and by multiple choice tests, with respect to *both* sex and school type.

7.1. The gender-linked discrepancies between ASAT and course scores that result in gender-linked biases in Tertiary Entrance scores reflect different processes for measuring educational properties. The discrepancies are not a statistical artefact of the process of aggregation.

7.2. ASAT scores, with or without Writing Task scores, are not sufficiently positively correlated with school-based assessments to justify rescaling course scores or general achievement measures without considering the necessity for their calibration to remove the gender-linked discrepancy between them and course scores. Calibrated scores have higher correlations with school-based scores. Other action to check on outlier scores may marginally affect the discrepancy measure.

8.1. The imprecision in a TE score is affected by choice of scaling procedure. Amongst procedures based on a one-factor model for the data, this imprecision is least when an Other Course Score procedure is used. The scaling parameters are model-unbiased when they are estimated by Method-of-Moments.

8.2. The major source of computational imprecision in TE score construction is associated with the use of ASAT scores. This imprecision can be considerably worsened by using the existing bivariate adjustment approach rather than the more direct estimation approach in a one-factor model for multivariate data.

9.1. The use of ASAT sub-scale scores as in the 1986 ACT scaling procedure for constructing TE scores contravenes Principle P 1.

9.2. Consideration should be given to redefining a TE score as the sum of a student's best 4.5 course scores, conditional on the inclusion of at least three Major course scores, where a Minor course has a weight of 0.5 in place of 0.6, and where a Total ASAT Score with components from the Quantitative and Verbal sub-scales and the Writing Task, may be included in this 4.5 course score total by regarding the score as having the weight of a Minor course.

10.1. Other Course Score scaling procedures have considerably smaller error mean squares than ASAT scaling procedures, and are therefore superior procedures in general.

11.1. On the basis of what is presently known, there are four steps required to construct a TE score as fairly as possible via a set of linear transformations of school-based scores, consistent with the Basic Assumptions and Principles for Constructing TE scores:

(1) Determination of scaling parameters via Method-of-Moment estimation using Other Course Scores as the basis of the scaling criterion variables, in conjunction with (2)–(4) below.

(2) Removal of the gender-linked bias in ASAT scores.

(3) Reduction of the effects of outlier scores (this may overlap with (2)).

(4) Fixing suitable weights for ASAT- (or whatever-) based reference scale scores in relation to non-statistically determined course scores, as for example with small groups.

12.1. The use of a particular scaling procedure can have considerable influence on the set of students meeting a TE score based selection criterion, particularly in the more selective groups. In comparison with the *OptOCSP*, the 1986–88 procedure is most noticeably discrepant, even before the removal of the gender-linked bias.

12.2. Using more than one reference scale implies that optimal subject choices can increase a student's TE score via statistical properties of a scaling procedure. Such choices may be contrary to "educationally desirable" curriculum construction.

12.3. Correlation coefficients do not usefully summarize differences in TE scores resulting from different scaling procedures.

12.4. No systematic bias effects are observable in the moderation group parameters computed via the Optimal Other Course Score Scaling Procedure.

13.1. The 1977–85 scaling procedure and the model described by equations (13.1)–(13.3) are not mutually consistent.

13.2. The major component of an ACT data set needing statistical scrutiny to ensure fair use of a statistical scaling procedure is the set of ASAT scores.

13.3. The existing scaling parameter equations (13.4)–(13.5) are not justified by Model 1 but are supported weakly by Model 2, i.e., a one-factor model. Method-of-Moment estimators with Other Course Score scaling gives a far more consistent fit to the one-factor model description of the data.

13.4. If multiple aggregates are constructed within restricted subsets of courses, the principles of Other Course score scaling using Method-of-Moment estimators and construction and use of reference scales as in Chapters 4–7, apply.

Chapter 1

Introduction and Summary

Mathematics and statistics:
languages for science and social science

The Project as Assigned

The task I have been assigned is part of a research programme continuing on from the 1986 Report *Making Admission to Higher Education Fairer*, referred to as *MATHEF* below. This report was prepared by a Review Committee set up in response to the following:

> "One major concern about the current method of calculation of Tertiary Entrance Scores in the Australian Capital Territory gave rise to this inquiry. It was a concern that the process is biased against females who are taught and assessed in groups that contain no males, [being] greatest for females in single-sex schools." (*MATHEF*, §1.1)

I have been asked by a 1986–87 Supervisory Committee[1] generally

> "to determine whether a [certain scaling procedure can] provide a statistically and educationally acceptable basis for continuing to produce a single aggregate Tertiary Entrance Score".

In particular I have been asked

(1) to state the theoretical basis of [a certain scaling] procedure . . . ; a clear statement of the algorithm . . . should be provided;

(2) [to supply data on the procedure to the ACT Schools Accrediting Agency] for checking purposes; and

(3) [to provide] data . . . comparing results [from different scaling procedures].

I responded to (2) in September 1988 by giving enough data from "a certain scaling procedure" to the ACT Schools Accrediting Agency to construct tables analogous to those that are published in each *Year 12 Study*. More informative examples of different Tertiary Entrance scores produced by different scaling procedures are given in Chapter 12, together with comparisons of their scaling parameters.

This introduction serves as a long summary of the report and is deliberately devoid of algebra (though not of technical ideas). It refers to the theory requested in (1), gives summaries of data analyses that are essential to provide some check that the theoretical results are applicable to ACT data, and applies the theory, completing the response to (1) and (3) to the extent that it is feasible.

A statement of principles and assumptions for constructing Tertiary Entrance scores in the ACT is given only in this chapter, as part of the general background.

The contents of this report are determined both by the nature of the data and the project, and by the implied constraints of the Supervisory Committee as to how they saw the matter through

[1] A short name for a three-member committee established jointly in late 1986 by the Australian National University, the Canberra College of Advanced Education, and the Australian Capital Territory Schools Authority, to supervise research into Tertiary Entrance Score Calculations, with Secretariat services provided by the ACT Schools Accrediting Agency.

1987: this view of the project in relation to the data is discussed briefly in an Appendix to this chapter.

Some General Background

The Tertiary Entrance (TE) score constructed in the Australian Capital Territory (ACT) is an example of an index that can be used to help select students for admission to higher education. The intention is that it reflect relative academic merit, as shown in academic performance at the upper secondary school level.

The central issue we address is how such an index can be constructed from the range of more specific "subject marks" or "course scores". These are indices of relative academic achievement. They can be produced by public examination, ability testing, any form of school-based assessment, or by combining any of these or similar types of indices. Phrased thus, the report is more than a narrow study of the relative merits of particular "mark scaling methods" for use in the ACT: a general framework enables us to see problems like those of the ACT in better perspective.

Across Australia, statistical procedures assist in producing aggregate scores that are "fairer" to all students and less amenable to artificial manipulation by choice of course or by other means. Their use can be traced to educationalists who were managing systems that incorporated widely administered public examinations and who recognized the possibility for such systems to be manipulated. Specifically, these procedures aim to ensure:

(i) that there should not be undue fluctuations in scores awarded to candidates in courses with large candidatures changing little in either numbers or nature; and

(ii) that a student's choice of subjects should not in itself lead to any advantage or disadvantage.

In this way, it was hoped to foster greater flexibility ("freedom of choice") for students in their curriculum construction. Further, an efficient centralized system of Tertiary Admission requires some automated selection. In this way statistical procedures have become an administrative necessity.

> **A major aim of statistical procedures is to translate the educational goals of consistency and equity into practice.**

Principles and Assumptions for Producing Tertiary Entrance Scores in the ACT

The following statement of principles and assumptions is based on a paper and subsequent discussion at a two-day seminar on aspects of Tertiary Entrance score construction held in Canberra in July 1988. The statement does not have any official standing: it is given essentially to provide some focus for our exposition. In particular, some like Masters & Beswick (1986) have queried A 1, while the sex bias problem contravenes P 4 because of a failure of A 5. "Short names" for the statements are given in parentheses.

Basic Assumptions

A 1. It is possible to measure relative general academic achievement on the basis of a student's work in Years 11 and 12 from their course scores. [TE scores are meaningful.]

A 2. Teachers can make valid and reliable judgments concerning the relative performance of their students. They can accurately represent these judgments by scores in the courses for which they are responsible. [School-based assessment is valid.]

Principles in Constructing TE Scores

P 1. The construction of a TE score should have a minimal effect on the curriculum choice of students, nor should students' curriculum choices *per se* affect their TE scores. [Independence of TE scores and subject choices.]

P 2. The system should preserve students' relative achievements as depicted by the teacher-determined scores. [Respect teachers' comparative judgments.]

P 3. The prime responsibility for detailed assessment of any course should lie with the college or school responsible for the conduct of that course. [No external exams.]

P 4. The TE scores of a group of students should not be influenced by anything other than their relative academic achievements. [No bias.]

Assumptions about ASAT Scores

A 3. Australian Scholastic Aptitude Test (ASAT) scores measure a factor of general scholastic aptitude. [What ASAT scores measure.]

A 4. General scholastic aptitude can be measured validly by a mix of questions from the areas of humanities, social science, mathematics, and the sciences. [ASAT scores reflect a mix of aptitudes.]

A 5. ASAT scores are sufficiently positively correlated with relative academic achievement to justify rescaling course scores. This "common scale" validly compares student achievement across all schools within the system. [A statistical procedure for TE scores can be based on ASAT scores.]

A Logical Basis for Scaling

Our development is governed by an adherence to Principles P 1–4 and acceptance (with some verification) of assumption A 1. On these bases, different starting points in Chapters 2 and 3 consistently lead to (A) and (B) below. The request "to determine . . . an educationally acceptable basis for a single aggregate Tertiary Entrance score" is equivalent[2] to developing a procedure consistent with a set of educationally agreed principles, so granted such acceptability of Principles P 1–4, (C) follows also.

This development is consistent with an otherwise unstated premiss, that

any valid scaling procedure should have a logical basis.

It is a premiss supported by Masters & Beswick (1986, §2.32):

> "A weakness of scaling procedures in use in every system in Australia is that they are not supervised by an explicit statistical model for score equating and aggregation."

The strength of such an approach is that it enables us to answer by implication, if not also directly, questions of the relative merits of different scaling procedures.

CONCLUSION 1.1. **If a composite measure of relative general academic achievement is to be constructed on the basis of several particular measures of achievement like exam. marks, then**

(A) it is statistically acceptable that it be an aggregate of scaled marks;

(B) the scaling should be determined internally by the set of marks in conjunction with the method-of-moment estimation procedure;

(C) an aggregate of marks determined via (B) is also educationally acceptable.

[2] "It is sometimes assumed in the educational field, if not others, that the adequacy of a system is measured by the ready comprehensibility of its technical details ... The Working Party has chosen the approach that what should be readily understood are the principles on which the system is based ... It has proved impossible to devise a system that both discharges some worthwhile principles and also has no technical aspects of greater complexity than can be explained to a general audience in a few words." (*Tertiary Entrance in Queensland: A Review* (1987), p.123).

What is a Scaling Procedure?

A *scaling procedure* is an algorithm that changes one collection of sets of exam. marks or course scores etc. into another collection of scores so as to preserve whatever orderings there are in the original sets, with the goal of estimating students' relative general achievements equitably. We consider only procedures in which the changes are effected by linear transformations. These have an appeal of simplicity as they are specified by determining what the means and standard deviations of the sets of scaled scores should be. The simplest of the ASAT scaling procedures states that these means and standard deviations for each set of scaled scores should coincide with those of the ASAT scores of the same sets of students; these sets refer to scores in different courses or groups of courses called *moderation groups* in the ACT. In the literature, justification for this procedure is often given via ideas of common scales for different groups. This justification is quite distinct from the motivation for having a scaling procedure in the first place, namely the "fair" construction of a measure of relative general academic achievement. Addressing this problem directly shows that the general setup of estimating general achievement reduces in a special case to this "common scale" approach, but that otherwise the two approaches are distinct. As for any logical connection between the motivation and the existing procedure, all that can be shown, in both theory and practice, is that it is a markedly inefficient method of estimating parameters to fit a one-factor model to such data as usually arise. In short, the existing procedure does a noticeably rough-and-ready job of meeting the aim.

First approaches

Chapters 2 and 3 give some basic analyses, introduced by the simplest of examples of "balanced" data sets for which there are only scale parameters to be determined (i.e., standard deviations for scaled course scores); this topic is much neglected in the literature. By using a direct representation for a typical data set, we show what *any* aggregate score constitutes, and use this to determine what constitutes a "fair" aggregate in practice when students study a subset of courses from a much larger pool of offered courses.

This "fair" aggregate essentially coincides with what is constructed by applying a Method-of-Moment (M-o-M) estimation procedure to the model in Daley & Seneta (1986); they advocated its use on the grounds that it involves making minimal assumptions about the data. The representation in Chapter 2 entails *no* modelling assumptions about the data, so it provides even more reason for using the model considered by Daley & Seneta together with M-o-M estimation.

Two bonuses come from this direct representation for a simplified data set. First, for a typical such data set, there is essentially only one set of scaling parameters that satisfy the equations that are easily set up, and second, the values of these parameters can be found by an iterative solution technique. From a mathematical point of view, these are highly satisfactory observations because hitherto the various so-called "iterative scaling procedures" in use in Australia have been known to converge to a solution merely on empirical grounds — we now know that, with a simplified data set, convergence necessarily occurs, and yields the essentially unique solution.

Rasch models are fashionable in educational measurement circles. The use of models akin to these in considering how course scores might be used to reflect relative academic merit, leads to formulating the same mathematical problem as is solved by the M-o-M estimation procedure. Because the scaling criteria for a given set of course scores comes from Other Course scores, and because the M-o-M parameter estimates have the critical property of unbiasedness amongst Other Course score estimates, we call it the Optimal Other Course score Scaling Procedure, and denote it *OptOCSP*.

Data Analyses

Much of our technical discussion and analyses centres on a description of the data set, aimed partly at checking the assumptions underlying the development in Chapters 2 and 3, and partly at repairing a serious deficiency of Masters & Beswick (1986), called M&B below. M&B failed to give any quantitative discussion of the multidimensional nature of course and ASAT scores, yet claimed that this multidimensionality is the root cause of the ACT's problems. They asserted without valid evidence that the sex-bias problem is a statistical artefact, being a manifestation of this alleged root cause.

For data of the type that arise with ASAT and ACT course scores, it is possible to describe almost all of the systematic variability by three measures: a general achievement/ability factor, a contrast factor between quantitative and verbal skills, and a contrast factor between ASAT and school-based course scores (see Chapters 6 and 10). In Chapter 6 we look for possible global dependency relations between these factors, for M&B alleged that such exist. We do not find evidence to support the claim.

The description in terms of three measures implies that such data sets may be fitted by a statistical model involving "true scores" that have a three-dimensional structure. In this model, TE scores are identified with the relative general achievement factor, modulo a much smaller perturbation produced by a fraction of the quantitative/verbal contrast factor. Use of this model, coupled with some knowledge of the sizes of the factors involved, allows us to consider a range of methods for estimating students' relative general achievements via measures like the existing Tertiary Entrance score (cf. also Daley & Seneta, 1986).

Data analyses in Chapter 8 show that *OptOCSP* is the scaling procedure with by far the least imprecision: thus it is arguably the most faithful to the school-based measures. Theoretical analyses in Daley (1988) point to the *OptOCSP* having model unbiased[3] scaling parameters. These two properties taken together lead us to deduce that an optimal scaling procedure respecting the essence of the existing constraints and principles is given by *OptOCSP*.

On the grounds of both biasedness and precision, the existing ASAT Scaling Procedure (*ASATSP*), first devised on an *ad hoc* basis in Queensland in the mid-70s and used in an essentially similar fashion since, is demonstrably inferior to the *OptOCSP*. This is because scaling parameters under *ASATSP* are no longer necessarily model unbiased, and both theoretically and empirically the imprecision of the resulting TE scores is much larger.

The properties of *OptOCSP* that make it consistent with the Basic Principles underlying the production of TE scores in the ACT are that

(i) it preserves the principle of independence of TE scores and subject choices;

(ii) it preserves teachers' comparative judgments of achievements;

(iii) it places prime responsibility for production of TE scores with the colleges much more faithfully than *ASATSP*; and

(iv) subject to the elimination of the sex bias problem existing between ASAT and school-based assessments, it shows neither advantage nor disadvantage for any groups of students.

[The sex bias problem arises from a gender-linked discrepancy between ASAT and course scores (Chapter 6). It is a phenomenon basically independent of the Scaling Procedure *per se*, and requires independent action to rectify it (see Chapter 7).]

[3] While not stated explicitly, so far we have essentially been describing how to judge the choice of procedures for determining the standard deviations. *Unbiasedness* here refers to the so-called scale parameter. Later we shall consider the more serious bias problems that can be associated with procedures for determining location parameters: the sex bias problem is of this latter type.

The *OptOCSP* is also consistent with Assumptions A 3–4 concerning ASAT scores, except that it is much less dependent on individual vagaries of ASAT scores, and its correlational properties are much better (Chapter 10). The *OptOCSP* is more resistant to manipulation than the *ASATSP*, in the sense that attempts to "play the system" so as to gain some relative advantage are accompanied by either penalties for wantonly trying to gain such advantage without treating the activity seriously, or else beneficial educational side-effects when students generally approach the whole range of their learning activities even-handedly as is assumed by any scaling procedure.

The relative general achievements of students within any college can be determined as Years 11 and 12 progress whenever any set of pseudo-course scores is available. The resulting indicators can be expected to change in time and when inter-college measures (presently, ASAT scores) are determined. The final within-college changes would be rather less than with predictions of comparable indicators under the *ASATSP*. If such "within-college TE scores" are used, students may become aware sooner of tertiary level courses likely to be open to them. On the other hand, making the *OptOCSP* algorithm available to colleges means that there would be available a facility to observe the effects on within-college TE scores of any manipulations of the basic course scores. Equally, it is open to colleges to write their own programmes to perform internal scaling (and I am aware of some institutions that do this, and not just at the Year 12 level).

Second Approaches

The major difficulty with the ACT and Queensland systems in terms of producing system-wide TE scores concerns how to establish inter-school comparability of any aggregate scores. Producing the within-college scores is not so difficult: between colleges is a much thornier problem, as the Accrediting Agency's records for 1978-88 bear testimony.

Since the argument of Chapters 2 and 3 points to fallacies associated with using bivariate equating procedures with multivariate data, the existing use of ASAT or any other reference scale score needs reappraisal. In Chapter 4 we start by finding an optimal mixture of ASAT sub-scale and Writing Task scores for use as a reference scale. For 1986, the system-wide optimal mixture of $Q : V : W = 47 : 18 : 35$ is consistent with the optimal mixtures formed within each college, and also within sex within each college. It is most desirable that the analyses be replicated with at least one more data set before any prescriptive determination of these weightings be made. Further, such replication should be a routine feature of the annual *Year 12 Study* so that any medium term shifts may be detected.

In principle, it is feasible to determine each year, when ASAT scores have become available and before the TE scores are finally produced, whether any marked shift from a pre-determined weighting of components of an ASAT Total score has occurred. However, I sense that the ACT community, while quite happy to accept recommendations on technical matters from advisors with a philosophical or dogmatic bent, may be less happy to accept them when empirically determined from analyses of long runs of similar data sets, even when there is at best limited quantitative understanding of what those analyses accomplish. We then consider how to relate a scale of one reference measure to the "scale" of a set of within-college measures of general achievement. It looks like a setting for a bivariate adjustment procedure, except that the ASAT and general achievement measures have errors of quite different orders of magnitude. The ultimate solution (Chapters 4 and 11) involves treating the optimal mixture of ASAT scores as scores from a course of value intermediate between that of a Minor and a Major course score.

RECOMMENDATION 1

(a) **As from 1989, a single-aggregate TE score should be constructed from course scores that have been scaled via the Optimal Other Course Score Scaling Procedure.**

(b) **Consideration should be given to making available to colleges, at certain intermediate times during Years 11 and 12, access to a pseudo-OptOCSP to enable pseudo within-college TE scores to be constructed, on the strict understanding that any pseudo-scores so produced are interim indicators and subject to moderate changes when finally determined.**

(c) **Work should be continued in conjunction with appropriate tertiary institutions on the informational content of a student's course and ASAT scores (and any other information that may be available) to determine whether any supplementary indicators may be constructed from these data to indicate whether that student's TE score on its own is inadequate as a general summary of the student's achievements.**

RECOMMENDATION 2

Analyses concerning Writing Task scores and their relation to ASAT subscale scores in determining an "ASAT Total Score" should be replicated, routinely monitored each year as part of the Year 12 Study, and some pre-determined weighting of $Q : V : W$ agreed on the basis of technical advice.

What we have not discussed so far is the use of reference scale scores to establish comparability across colleges. This is the role of ASAT scores as used in the ACT, WA and Queensland (in much of the discussion, "ASAT" and "reference scale" are synonymous). Two parameters are entailed, affecting the location and scale of scores respectively. We assume that only the raw indices that are currently combined into the ACT TE score are available. Then the existing procedure for constructing TE scores is shoddy and should be replaced by one that is more akin to those in use by most other Certification bodies in Australasia. Such a procedure is described in Chapter 11.

As a by-product of noting discrepancies between reference scale scores and school-based course scores, we also find more evidence indicating that the gender-linked bias problem that has plagued ACT TE scores since first issued in 1977, has not been solved. On the evidence it is unlikely to be solved unless either more external examining is restored, or statistical adjustments are made to the scores from the ASAT test. This is because the bias is a product of different educational measurement methods. It emerges as a gender-linked discrepancy between these ASAT scores and the school-based assessments that are proclaimed to be the norm for certification in the ACT. We note in Chapter 7 some of the deficiencies of earlier reports in documenting the matter.

The majority view of the Review Committee that prepared *MATHEF* was an hypothesis proposed in Masters and Beswick (1986) that

> "the major source of current problems with Tertairy Entrance Scores in the Australian Capital Territory is the multidimensionality of the scores from different courses which are combined to form the Tertiary Entrance Score." (*MATHEF*, §7.1)

MATHEF recommended certain changes that admitted the existence of a sex bias. Since I have been asked to look at the possibility of maintaining the use of a single aggregate score, when the operation of constructing such scores is supposedly a cause of the bias, it is relevant to check whether the majority view remains tenable in the light of the further information that has emerged since *MATHEF* was written in June 1986. Indeed, notwithstanding Recommendation 1 of *MATHEF* that

> "there should be no adjustment to Australian Scholastic Aptitude Test scores by calibration on the basis of students' sex",

the Supervisory Committee has asked specifically that I calculate

> "the calibration which would be required were a statistical adjustment to be made to the scaling criterion on the basis of sex."

This is given in Table 12.5.

It follows from a wide-ranging review (Daley, 1989) that the Committee's majority view just quoted was not well founded. Indeed, the only evidence produced by M&B in support of their hypothesis is fundamentally flawed, while calculations which give deliberate emphasis to any effects of multidimensionality on the sex bias measure fail to produce the hypothesized consequence. The conclusion of all this is that M&B's hypothesis is vacuous. Direct use of educational measurement data reinforces this conclusion.

In contrast, the more recent evidence shows that what the Review Committee regarded as a second order effect in producing the bias, is in fact the dominant source: the bias arises primarily from a gender link in the different skills and attributes reflected in Australian Scholastic Aptitude Test multiple-choice based scores on the one hand and school-based assessments on the other.

Three obvious avenues for the ACT Schools Accrediting Agency to consider for use in removing the sex bias are as follows:

(1) a statistical calibration of ASAT scores as discussed in *MATHEF*;

(2) increasing the weighting of the Writing Task in the so-called ASAT Total score;

(3) the use of some common external written assessment in the mathematics/science area, more specifically norm referenced and curriculum oriented than the Writing Task, so as to secure a measure that more closely reflects school-based assessment of achievement.

I give some discussion of the merits of each of these approaches and a fourth less obvious route to enable the ACT Schools Accrediting Agency to reach its own conclusions on how the problem may be dealt with effectively.

RECOMMENDATION 3

> **Appropriate action should be taken forthwith to eliminate gender-linked biases in TE scores.** (*MATHEF*, §1.1).

From a personal point of view, and also from an historical prespective, technical problems relating to ACT course and TE scores have been treated with greatest competency and expediency when the technical workers concerned have interacted in a forum of about half a dozen interested people to assist in a "multidisciplinary" exercise. An appropriate model for technical or scientific work is not the adversarial one involving "expert witnesses" but rather the cooperative evaluatory one of "friends of the court". A major reason for disagreement in 1985 between the ACT Schools Accrediting Agency and its Technical Advisory Committee was the unwillingness of the Agency to allow its Technical Committee any real participation in its decision making processes which tended to be political even on technical matters. It is unreasonable to expect the Agency's Secretariat to maintain in its employ for a sufficiently long time individuals with all the skills required to make apposite assessments of the range of technical matters that can arise.

Comments in the Appendix to this chapter give further support to the following.

RECOMMENDATION 4

> **A Technical Advisory Committee to the ACT Schools Accrediting Agency should be revived, with responsibility for providing advice on technical matters relating to the construction of TE and course scores. Its membership should be longer term and not necessarily representative of interest groups. It should be responsive to but not constrained by the Agency. Where the Agency is unwilling to accept advice from the Committee on technical matters, it should meet jointly with the Committee to reach decisions by consensus.**

Appendix to Chapter 1

The Nature of the Project

The formally assigned task, as the 1986–87 Supervisory Committee saw it in an enquiry to me in December 1986, arose from the first part of Recommendation 8 of *MATHEF*, namely

> "Further investigations should be undertaken of the method of [moment estimation procedure for] scaling course [scores] against other course [scores] within colleges and against the Australian Scholastic Aptitude Test [scores] between colleges in order to determine whether an improved single aggregate may provide a sufficient and unbiased account of student performances."

In response to that enquiry, a project covering at least three items was outlined:

(i) an examination of what Masters and Beswick called the multidimensional nature of the data set consisting of course and ASAT scores, because in spite of this idea being fundamental to their arguments, they did not investigate it;

(ii) the preparation and checking of computer programmes to process the data set using Method-of-Moment estimators in an Other Course Score scaling procedure, along with output of information indicating its precision; and

(iii) developing suitable procedures that should effectively eliminate the effects of the gross differences between ASAT and Tertiary Entrance scores because they affect the "fairness" of TE scores no matter what scaling procedure[4] is used.

The Supervisory Committee amended considerably this view of how to address Recommendation 8 by cutting the first and third items, and asking that this report

> "specifically state the theoretical basis of the method of moments approach to scaling, addressing the concerns expressed by the [Review] Committee which prepared *MATHEF*."

The response to this request for a theoretical statement is almost certainly met more fully than anticipated[5] by Chapters 2–4 and Daley (1988). The concerns are met by implication: the approaches through data representations in Chapter 2, modelling in Chapter 3 (and Daley & Seneta (1986) before that), unbiasedness and precision considerations of Daley (1988) and Chapter 8, regression analyses in Chapter 10, and the correlations listed in Tables 12 of the *Year 12 Study*, all point to the superiority of Other Course Score scaling procedures using a single reference scale over the present bivariate adjustment procedure, whether one or several reference scales are used. Much of this depends on the structure of the data set and data analyses, which are needed to validate the use of procedures predicated on the positivity assumptions of Chapter 2 or modelling assumptions of Chapter 3, yet item (i) was specifically not requested.

Similarly, the Committee's request that

> "a clear statement of the algorithm to be used should be provided"

[4] These gross differences affect the fairness of *any* Tertiary Admissions Index that is consistent with all but A 5 of the Principles and Assumptions for constructing TE scores in the ACT. The effects need not be visible without analysis, in which case they can be overlooked or [worse] ignored. For example, the gender-linked bias effect has been noted ever since ACT TE scores were introduced.

[5] The interim report was prepared before taking this request into consideration: essentially Chapter 2 and its approach comprise the specific response. Chapter 4 arises in part to indicate that there is a range of apparently different mixtures of the ASAT sub-scale and Writing Task scores that do not differ significantly, a facet not brought out in similar analyses by Morgan & McGaw (1988).

cannot be met within the constraints it imposed. For a start, the gender-linked bias between ASAT and course scores should be eliminated. Then, we should reduce further the effects of excessive deviations from the supposedly sufficiently close relation between ASAT and course scores that has in the past been assumed to be adequate to justify the existing scaling procedure. As a compromise, I have exceeded the constraints of the Supervisory Committee and indulged in some *ad hoc* computational work.

All that is being noted here is that,

> **within the circumscriptive terms of the present commission, specification of an algorithm that would be both fair to all groups of students and ready for implementation is not technically feasible.**

In Chapters 4 and 11 there is shown how to construct an optimal reference scale score and an algorithm for scaling via Other Course Scores using Method-of-Moment estimation is specified. Calibration to correct for gender-linked biases has not been incorporated, nor are any details included of how to cope with outlier scores.

The Committee also sought to specify certain summative information for the various scaling procedures from which it is all too easy to make inappropriate comparisons between scaling procedures, rather than asking more basic questions such as whether a scaling procedure is justifiable, and if so is there an optimal way of reflecting that justification? Valid comparisons can be made between scaling procedures, but mostly by criteria different from those that the Supervisory Committee specified.

There are specific examples of both existing and rival TE scores, encoded to obscure individual student identity without essential loss of information. These serve to illustrate how the range of TE score behaviour can be summarized by more suitable statistical criteria. The various moderation group parameters are also compared for the different scaling procedures.

In this report data based on the whole of the ACT population seeking a TE score are sometimes given without identifying the colleges concerned: this has been done in an attempt to focus attention on the content of the data without the distraction of knowing from what college the data come.

The work done at the time of submitting an interim report largely reflected the requests of the Supervisory Committee indicated to me when I was provided with a data tape early in August 1987. It is since that time that the requests to "state the theoretical basis [and] the algorithm" have been added, without any consultation as to the feasibility of describing a *fair* algorithm (as noted, this last is incompatible with the selection of work sought and the properties of the data set).

Chapter 2

Representations of Course Scores and The Construction of Aggregates

"*A weakness of scaling procedures in use in every system in Australia is that they are not supervised by an explicit statistical model for score equating and aggregation.*" (Masters & Beswick, 1986, §2.32)

What an Aggregate Score Represents

Adding up "marks" is an operation with which most of us are familiar from contexts where students have undergone assessment of more than one task, starting from the simplest examples of spelling or arithmetic tests. This chapter starts by looking at a slightly more complex setup which is "simple" relative even to a public examination system such as in New South Wales or Victoria. We choose this approach because by describing the simple setup we shall demonstrate the core of the so-called scaling problem that arises in aggregation, by which we mean the operation of adding up marks that define rankings, when the aim of the operation is the construction of an overall, composite ranking.

As a start suppose that all students who are being assessed follow the same curriculum and have been assessed in exactly the same set of courses. In each course the set of scores (or, marks) represents a ranking of the various students' achievements as assessed. Adding up these marks over all the courses and regarding the sum as defining a set of scores for a new ranking, is equivalent to defining the new ranking to be that combination of the various constituent elements of the component rankings in exactly whatever their representations are in the original scores.

Consider the very simplest such example, of just two sets of scores, in spelling and arithmetic say. Suppose these scores are "marks out of 100" (for example, the number of correctly spelt words out of 100 and the number of correct elementary sums out of 100); then the sum of the two scores is a well-defined mix of rankings. Consider what happens if the tests are at about an appropriate level of difficulty in arithmetic but too easy in spelling, so that there will be smaller differences in marks for average and above-average spelling performance but larger differences for below average. Then for most students, the aggregate will be influenced more by the spread of arithmetic scores than spelling, the exception being the minority group with appreciably lower than average spelling scores. It follows that for (say) the upper 50% of students, the aggregate score will reflect arithmetic rankings much more than spelling rankings.

In this example, and in any context where all students follow the same curriculum, the simple addition of marks can thus reflect different balances of attainments, even though on the surface it is being defined "fairly" as the sum of (say) the same number of "marks out of 100". It was just this sort of unintended variation in the "definition" of an aggregate, as examination papers changed from one year to another, that made educational administrators look for ways of preserving some constancy over the years and thereby eliminate a major source of unintended variability that may otherwise result (as *e.g.* "the English paper in my year was so easy we all got good marks", little realizing that in the "class averages" for the year, while the overall aggregate marks may have been raised as a consequence, English marks would have played a lesser role in terms of determining any ranking; this example should not be quoted out of its purely illustrative context).

Return now to the general situation, where all students follow a common curriculum and have scores in exactly the same set of courses $j = 1, \ldots, n$. Assume that student i has a score X_{ij} in course j, and that the relative academic merit of students is defined to be the ranking determined by the sums

$$T_i = X_{i1} + \cdots + X_{in}\,. \tag{2.1}$$

Given any such multivariate data set $\mathcal{X} \equiv \{X_{ij} : j = 1, \ldots, n;\ i = 1, \ldots, N\}$ for which it is assumed without loss of generality that $\sum_i X_{ij} = 0$ and $N \geq 2$ (usually, N is rather larger than n), there exists the principal component representation

$$X_i = LU_i \tag{2.2}$$

for the n-vectors $X_i \equiv (X_{i1}\ \cdots\ X_{in})'$ in terms of vectors $U_i \equiv (U_{i1}\ \cdots\ U_{in})'$ which have orthogonal components in the sense that

$$\sum_i U_{ij}U_{ik} = \delta_{jk}\lambda_j \tag{2.3}$$

for constants $\{\lambda_j : j = 1, \ldots, n\}$ which are the eigenvalues of the covariance matrix

$$\Sigma_X \equiv N^{-1}\sum_{i=1}^{N} X_i X_i' \tag{2.4}$$

and L is the matrix of associated eigenvectors (see e.g. Jolliffe (1986) or §8g.2 of Rao (1973)). In terms of this representation it follows that the sums T_i are expressible as

$$T_i = \sum_{j=1}^{n} X_{ij} = \sum_{j=1}^{n}\sum_{k=1}^{n} \ell_{jk}U_{ik} = \sum_{k=1}^{n} \ell_{\cdot k}U_{ik} \tag{2.5}$$

where for the column sums (= sum of elements in eigenvectors) we write

$$\ell_{\cdot k} \equiv \sum_{j=1}^{n} \ell_{jk} = \mathbf{e}'\ell_k \tag{2.6}$$

with $\mathbf{e} \equiv (1\ \cdots\ 1)'$. For any data set $\mathcal{X}$ as above, the covariance matrix Σ_X is positive definite, so all its eigenvalues λ_j are positive and they can be labelled as in

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0, \tag{2.7}$$

(and with data sets it is almost invariably the case that these eigenvalues are all distinct, so we shall assume they are, and thus that strict inequality holds). Finally, because L is orthogonal, (2.2) implies that $U_i = L'X_i$.

For data sets like $\mathcal{X}$ it is usually the case that the covariance matrix Σ_X has all its entries positive, and fairly substantially so. Such positivity implies that

(a) $\lambda_1 > \lambda_2$, and

(b) all the components of the eigenvector $\ell_1 \equiv (\ell_{j1})$ associated with λ_1 are positive.

Typically, λ_1 is the largest of the eigenvalues by an order of magnitude, while the positivity of the covariances is such that, when the variances of $\{X_{ij}\}$ are about the same order for each j, the components of ℓ_1 are all about the same size, and thus are approximately equal to $1/\sqrt{n}$. The orthogonality of distinct eigenvectors then implies that, except for $\ell_{\cdot 1}$, the sums of components

$$\ell_{\cdot k} \approx 0 \quad (k = 2, \ldots, n). \tag{2.8}$$

The implications of these properties for the sums $\{T_i\}$ are that in writing now

$$T_i = \ell_{\cdot 1}[U_{i1} + \textstyle\sum_{k=2}^{n}(\ell_{\cdot k}/\ell_{\cdot 1})U_{ik}], \tag{2.9}$$

the variance λ_1 of the set $\{U_{i1}\}$ of first components is largest by an order of magnitude, while the coefficients $\ell_{\cdot k}/\ell_{\cdot 1}$ are approximately zero. In practical terms this then means that, while the aggregate T_i is by definition composed here of a certain differential mix of the base scores X_{ij}, it is in reality dominated by the multiple $\ell_{\cdot 1}U_{i1}$ of the first principal component which has rather larger variability than other components and a substantially larger multiplier. We summarize as follows.

Conclusion 2.1. **A ranking determined by aggregates from data sets such as $\mathcal{X}$ is basically a ranking determined by the first principal components $\{U_{i1}\}$ in the principal component representation**.

Example 2.1. For the purpose of demonstrating principal component analyses rather than displaying a particular data set, we present matrices and eigenvectors for a set of data from an ACT college in 1986. The data consist of ASAT sub-scale and Writing Task scores and (scaled) English and Mathematics scores, for all students with those five scores (*c.* 90% of students). Three sets of tables related to this data set are shown. The first is of the covariance matrix of the scores $\{X_{ij}\}$ as issued, rescaled for convenience by division by $625 = 25^2$, the second is of the matrix of correlation coefficients, while the third is for the data rescaled so as to have all entries in the eigenvector associated with the first principal component equal to each other. The sums $\ell_{\cdot j}\sqrt{\lambda_j}$ are shown for the first two matrices; in the third case the sums are zero by orthogonality with the first eigenvector. There are other examples in this report (e.g. Table 5.3).

Example 2.2. Consider an $n \times n$ matrix whose diagonal elements are 1 and all other elements equal a for some a in $0 < a < 1$. Its eigenvalues are $1 + (n-1)a, 1-a, \ldots, 1-a$. For example, when $a = 0.5$ and $n = 5$ the eigenvalues are 3.0, 0.5, 0.5, 0.5, 0.5.

Two points are worth stressing here. First, because the scores of the set $\mathcal{X}$ represent relative academic achievement on n scales with $n \geq 2$, there is a whole family of "aggregate" measures of relative achievement which it is possible to construct as the set of aggregates of positive multiples of the elements X_{ij} of $\mathcal{X}$. In general, these relative achievement measures will change as the multipliers change, emphasizing that there is an implicit arbitrariness about the n scales of the scores in the first place. Second, Conclusion 2.1 is based on a mathematical representation which necessarily exists so soon as the data set $\mathcal{X}$ has positive correlations.

Remark 2.1. **Conclusion 2.1 does not depend on any of the statistical modelling assumptions to be discussed in Chapter 3.**

TABLE 2.1

Covariance Matrices of ASAT component, English, and Mathematics scores, with Eigenvalues and Eigenvectors ("loadings")

(a) Scores as used in 1986

ASAT Q	0.734				
ASAT V	0.579	1.076			
Writing Task	0.411	0.373	0.959		
English	0.494	0.614	0.679	1.058	
Mathematics	0.451	0.368	0.345	0.539	0.663
Eigenvalues λ_j	2.886	0.691	0.431	0.329	0.154
Eigenvectors* $\times \sqrt{\lambda_j}$:					
ASAT Q	0.696	-0.213	0.184	-0.349	-0.221
ASAT V	0.809	-0.560	-0.300	0.057	0.125
Writing Task	0.736	0.532	-0.255	-0.239	0.110
English	0.913	0.220	0.002	0.381	-0.175
Mathematics	0.611	0.013	0.492	0.040	0.216
Sums	3.765	-0.008	0.123	-0.110	0.055

(b) Correlation Matrix

ASAT Q	1.000				
ASAT V	0.606	1.000			
Writing Task	0.419	0.408	1.000		
English	0.504	0.601	0.645	1.000	
Mathematics	0.647	0.285	0.300	0.497	1.000
Eigenvalues $\{\lambda_j\}$	2.981	0.829	0.640	0.374	0.176
Eigenvectors $\times\sqrt{\lambda_j}$:					
ASAT Q	0.828	-0.347	-0.195	-0.328	-0.221
ASAT V	0.759	0.203	-0.584	0.071	0.192
Writing Task	0.714	0.484	0.396	-0.297	0.108
English	0.849	0.252	0.141	0.393	-0.205
Mathematics	0.700	-0.609	0.289	0.137	0.191
Sums	3.850	-0.017	0.147	-0.024	0.162

(c) Rescaled Scores**

y_1	0.835				
y_2	0.598	1.007			
y_3	0.457	0.377	1.045		
y_4	0.454	0.512	0.611	0.786	
y_5	0.572	0.424	0.429	0.553	0.939
Eigenvalues $\{\lambda_j\}$	2.917	0.696	0.529	0.312	0.158
Eigenvectors $\times\sqrt{\lambda_j}$:					
y_1	0.764	0.237	-0.067	-0.400	-0.178
y_2	0.764	0.485	0.391	0.124	0.140
y_3	0.764	-0.611	0.228	-0.141	0.128
y_4	0.764	-0.169	0.013	0.332	-0.253
y_5	0.764	0.058	-0.565	0.084	0.164

*In the language of factor analysis, the quantities $\ell_{jk}\sqrt{\lambda_j}$ are called loadings because they reflect the contributions or weights of the unit variance standardized factor scores $U_{ij}\ /\sqrt{\lambda_j}$ on the scores X_{ik}.

** Write $x_1, \ldots, x_5$ for scores underlying the matrix in (a). Rescale these to produce scores $y_1, \ldots, y_5$ by $y_1 = 1.067x_1$, $y_2 = 0.967x_2$, $y_3 = 1.044x_3$, $y_4 = 0.862x_4$, $y_5 = 1.190x_5$.

When is an Aggregate a "Fair" Aggregate?

Conclusion 2.1 leads immediately to asking about the composition of this principal component. We noted below (2.7) that $U_i = L'X_i$, from which it follows that

$$U_{i1} = \ell_{11}X_{i1} + \cdots + \ell_{n1}X_{in} \approx (\ell_{\cdot 1}/n)T_i\,. \tag{2.10}$$

Observe that the scores $Z_{ij} \equiv X_{ij}/\ell_{j1}$ have the representation

$$Z_{ij} = U_{i1} + \sum_{k=2}^{n}(\ell_{jk}/\ell_{j1})U_{ik}\,, \tag{2.11}$$

or in matrix notation, $Z_i = D_1 X_i$ where D_1 is the diagonal matrix $\mathrm{diag}(\{1/\ell_{j1}\})$ which is close to the multiple $\sqrt{n}\,I$ of the identity matrix. Suppose that it is prescribed that each course should contribute equally to the ranking determined by the aggregate of scores. Then since the ranking effectively reflects U_{i1}, it follows from the representations that the various sets of course scores influence the ranking quantity "equally" when the scores that are used are like Z_{ij} in having the same coefficient of U_{i1}, rather than variable coefficients as for the original scores X_{ij}.

Let us indulge in some heuristic mathematics. Consider matrices like the covariance matrices Σ_X observed with data sets like $\mathcal{X}$, specifically, positive definite with all entries positive and bounded away from zero. The vectors $Y_i = DX_i$ for any diagonal matrix D that is close to the identity matrix I, yield a covariance matrix Σ_Y say that is close to Σ_X, with eigenvalues and eigenvectors close to those of Σ_X. Denote its principal components by V_i (cf. (2.2)). Recalling Z_i at (2.11) prompts us to seek a diagonal matrix D_0 such that when we form the aggregate T_i defined as the sum $\mathbf{e}'Y_i$ of the components of the scaled course score vector Y_i, equivalently, a linear combination of the principal components V_i, only the first principal component appears in the representation $T_i = \mathbf{e}'LV_i$ as the sums of the eigenvectors associated with all other eigenvalues, i.e., their scalar products with $\mathbf{e}$, vanish because $\mathbf{e}/\sqrt{n}$ is an eigenvector and eigenvectors are mutually orthogonal. Thus, the approximations in the paragraph around (2.8) all become equalities.

Why should this matter at all? Is it not the case that one representation is about as good as any other if the sums of coefficients are all close enough to 0? Suppose that the aggregate, instead of being constituted as the sum of all scores, is constructed out of a subset of just m of the scores. Then using the scores $\{Y_{ij}\}$ we would have

$$T_i(m) = (m/\sqrt{n})V_{i1} + \sum_{k=2}^{n}\left[\sum_{j \text{ in subset}} \ell_{jk}\right]V_{ik}\,. \tag{2.12}$$

For example, if the scores were chosen as a student's "best m scores", the first term would remain unchanged, while the others would contribute as a mix of different subsets. By way of contrast, if the original scores X_{ij} were used rather than Y_{ij}, then the first principal component would have variable coefficients $\sum_{\text{best } m \text{ subset}} \ell_{j1}$, and students with scores in courses with smaller coefficients ℓ_{j1} in their "best" courses would be regressed towards the mean in comparison with other students. For example, McGaw (1987) makes this point in the context of a one-factor model.

In other words, amongst all the linear combinations that can be constructed from scores $\mathcal{X}$, we identify the combination which is best in accord with the "curriculum parity" principle that "each subject is counted so as to have the same influence in determining overall general achievement", as that combination that leads to a set of scores with representations in which the first principal components of the scores have the same coefficient for each course score. This parity principle must hold if P 1 is to hold. We state this more formally.

CONCLUSION 2.2. **When all students follow a common curriculum, the construction of a "best m subset" aggregate from course scores $\mathcal{X} \equiv \{X_{ij}\}$ most faithfully represents the principle P 1 that students' curriculum choices should not affect their TE scores when student i's scores $X_i = (X_{i1} \cdots X_{in})'$ are rescaled to scores $Y_i = (Y_{i1} \cdots Y_{in})'$ via relations**

$$Y_{ij} = \beta_j X_{ij}$$

for constants β_j such that the matrix $B\Sigma_X B$ with $B = \mathrm{diag}(\{\beta_j\})$ and Σ_X the covariance matrix of $\mathcal{X}$ has $\mathbf{e}/\sqrt{n} \equiv (1/\sqrt{n} \cdots 1/\sqrt{n})'$ for the eigenvector associated with its largest eigenvalue.

For the sake of consistency with existing literature we have written B here for the diagonal matrix denoted D_0 earlier. Note that B is at best determined only up to a scalar constant: we could for example require that $\det(B) = 1$.

Conclusion 2.2 embraces two mathematical problems. First, given any positive definite matrix Σ_X all of whose entries are positive, does there necessarily exist a diagonal matrix B with positive elements such that $B\Sigma_X B$ has an eigenvector proportional to $\mathbf{e}$? And if such a diagonal matrix B exists, is it essentially unique (i.e., up to constant scalar multiplier)? Both problems have been solved, and affirmatively: Sinkhorn (1964) proved a general result and noted an abstract of Marcus and Newman (1961) that gives the special case we need. See the Appendix to this chapter.

The rather more practical problem of determining the vector $\beta = (\beta_1 \cdots \beta_n)'$ for $B = \mathrm{diag}(\beta)$ now arises, the direct approach being to find a set of equations which its components satisfy. Recall the representation (cf. (2.2), (2.11) and Conclusion 2.1)

$$\beta_j X_{ij} = Y_{ij} = (1/\sqrt{n})V_{i1} + \textstyle\sum_{k=2}^{n} \ell_{jk} V_{ik} \,. \tag{2.13}$$

Because each vector $\ell_k = (\ell_{1k} \cdots \ell_{nk})'$ is orthogonal to the vector $\mathbf{e}$, summing on j in the right-hand side of (2.13) leads to

$$T_i/n = (1/\sqrt{n})V_{i1} \equiv \bar{y}_i \,, \tag{2.14}$$

and thus, using $\sum_i X_{ij} = 0 = \sum_i Y_{ij}$ for $j = 1, \ldots, n$,

$$\frac{1}{N}\sum_{i=1}^{N} Y_{ij}\bar{y}_i = \beta_j \,\mathrm{cov}(X_{ij}, \bar{y}_i) = \frac{1}{N}\sum_{i=1}^{N} V_{i1}\bar{y}_i = \mathrm{var}(\bar{y}_i) \quad (j = 1, \ldots, n). \tag{2.15}$$

These equations can be rewritten as

$$\beta_j = \frac{\mathrm{var}(\bar{y}_i)}{\mathrm{cov}(X_{ij}, \bar{y}_i)} \quad (j = 1, \ldots, n), \tag{2.16}$$

but they are not explicit expressions for β because $\bar{y}_i$ also depends on β. Intuitively, we would expect $\bar{x}_i \equiv (X_i'\mathbf{e})/n$ to be a reasonable first approximation to $\bar{y}_i$, and this is indeed the case. In an appendix to this chapter we show that equations (2.16) have exactly one solution that is consistent with Conclusion 2.2, and specify an algorithm involving an iterative process that converges and determines the solution. In fact, this iterative technique coincides with Sinkhorn's existence proof for B.

CONCLUSION 2.3. **The scaling constants β_j of Conclusion 2.2 are determined uniquely apart from a multiplicative constant, and there is a convergent iterative procedure to determine them.**

What do Scores Represent? Another Reason for using Aggregates

We started our discussion without asking an even more basic question as to what a score X_{ij} represents. Typically, it will be constructed from a collection of other scores (marks on separate questions in a test paper, or aggregation of marks from separate papers, etc.): how can we interpret such scores? The time when I was a pure novice in the area of "the statistics of examination marks" coincided by chance with my preparation of the 1983 Belz lecture to the Victorian Branch of the Statistical Society of Australia, having chosen the topic of ranking with special reference to the pecking order problem. It was an opportunity to contemplate precisely the question just given, and there came the suggestion of regarding marks $X_{i'j}$ and $X_{i''j}$ of individuals i' and i'' in course j as having the following local interpretation. Suppose all the scores were redetermined from scratch (e.g., different teachers, different class apart from the two students i' and i'' say); then interpret X_{ij}'s as the indicator

$$\frac{\Pr\{\text{student } i' \text{ ranked higher than } i''\}}{\Pr\{\text{student } i' \text{ ranked lower than } i''\}} \approx \exp\left(\frac{X_{i'j} - X_{i''j}}{\tau_j}\right) \tag{2.17}$$

for some scale factor τ_j that here represents a measure of the precision with which the measures $X_{\cdot j}$ enable us able to determine the relative odds as just shown. I little realized that this is similar to what follows from the Rasch model for item scores in multiple choice tests. This is not surprising: scores X_{ij} can be used to reflect a ranking of individuals, whether relative to each other or relative to some hypothetical standard, and all that is being given is a parametric model to describe this ranking.

We now have the mathematical problem of providing a framework for the combination of the scores into a representation of a composite ranking. At first sight this appears (and is) more difficult in the case of the relative odds representation just given than in the case of the linear factor model of Chapter 3. An answer to the problem should allow for prescriptive educational definitions as to "how much weight is to be given to each course", such decisions being constructive decisions based on educational considerations. In fact, in the case that all courses are weighted equally, we can be led to the same mathematical problem, as we now show. For this purpose, replace i' by i and i'' by some "standard" individual i_0 with score ξ_j in course j. Now define a "conglomerate relative odds function" for individual j as the product[6] of the component relative odds, and rank individuals according to this conglomerate function. Taking logarithms in no way affects this ranking, i.e., it is the same as using

$$\log(\text{conglomerate odds for } i) = \sum_j \frac{X_{ij} - \xi_j}{\tau_j}\,, \tag{2.18}$$

for student i, summation taking place over appropriate subsets of courses j. But this just states that the ranking is now being determined by an aggregate score.

CONCLUSION 2.4. **Constructing a general achievement index from independence assumptions and a quasi-Rasch model that uses course scores as parameters leads to the same starting point as the beginning of Chapter 2.**

Brief Review and Preview

The discussion leading to Conclusion 2.1 makes explicit the following: the fact that any aggregate is essentially an estimator of V_{i1} with the type of data set $\mathcal{X}$ as usually arises, is a consequence

[6] When probabilities of outcomes in different courses are independent, this product of odds ratios can be interpreted as the ratio Pr{student i ranked higher than i_0 in all courses}/Pr{student i ranked lower than i_0 in all courses}.

of the nature of the data set and not a result of an assumption being built into a model used to assist us in deciding how V_{i1} should be estimated. The later discussion leading to Conclusion 2.2 is prompted in part by an attempt in Appendices 1 and 2 of *Tertiary Entrance in Queensland: A Review* (1987) "to analyse the problem of how best to scale school-based assessments for the purpose of calculating [a relative general achievement measure]". This discussion has had some pleasing consequences: it has provided added justification for using the Method-of-Moment estimation procedure, and it has been related to work showing the convergence of "iterative scaling" for such an estimation procedure.

We turn in Chapter 3 to linking modelling approaches to the analysis of this chapter. These approaches are needed because, as any reader familiar with student curriculum patterns at the upper secondary level may be aware, the real situation is certainly more complex than the simplified curriculum model considered so far: it is usually the case that students can and do choose a subset of courses from the total range of courses offered.

A further problem, peculiar to systems such as Queensland or the Australian Capital Territory relying (almost) totally on school-based assessment, is how such scores as $\{V_{i1}\}$ produced within a school can be used to yield aggregates that correspond to an equitable ranking across all schools. This is considered in Chapter 4.

Mathematical Appendix to Chapter 2

Determination of the Scale Parameters: So-called Iterative Scaling Procedures

As a mathematical interlude, we sketch here some of the detail concerning the vector β of scaling parameters $\{\beta_j\}$ in Conclusions 2.2 and 2.3. In popular accounts of "scaling procedures", what we describe is one particular "iterative scaling procedure", which is a misleading name because in reality it is only the equation-solving technique that is iterative; the scaling procedure is an "other course score" procedure. For the sake of being more self-contained, we recall some of the assumptions and notation from the main body of the chapter.

Suppose given a data set $\mathcal{X} \equiv \{X_i : i = 1, \ldots, N\}$ of n-vectors of course scores for which the covariance matrix Σ_X has all entries positive (cf. (2.1)). The argument before Conclusion 2.2 leads to seeking a diagonal matrix $B = \operatorname{diag}(\beta)$ such that the covariance matrix Σ of $\mathcal{Y} \equiv \{Y_i\} \equiv \{BX_i\}$ is expressible as

$$\Sigma = L\Lambda L' \tag{2A.1}$$

in which $\Lambda = \operatorname{diag}(\{\lambda_1, \ldots, \lambda_n\})$ is the diagonal matrix of eigenvalues ordered downwards from the largest λ_1 as at (2.7), and L is the orthogonal matrix of eigenvectors $\{\ell_j\}$ of which the first is

$$\ell_1 = \mathbf{e}/\sqrt{n} \quad \text{where} \quad \mathbf{e} = (1 \ \cdots \ 1)'; \tag{2A.2}$$

uniqueness can be determined by requiring either

$$\det(B) = 1 \quad \text{or} \quad \beta_1 = 1. \tag{2A.3}$$

What follows is based on §2.6 of Seneta (1981) and Sinkhorn (1964) where more general details can be found.

Sinkhorn's Theorem 1 states that to a given strictly positive $n \times n$ matrix A there corresponds exactly one doubly stochastic matrix A_1 which can be expressed in the form $A_1 = RAC$ for diagonal matrices R and C with positive diagonals, with R and C themselves being unique up to a scalar factor (these diagonal matrices multiply Rows and Columns of A respectively). Express the doubly stochastic property of A_1 in the form $A_1\mathbf{e} = \mathbf{e}$ and $\mathbf{e}'A_1 = \mathbf{e}'$, i.e., $\mathbf{e}$ is both a left- and right-eigenvector for A_1. In our case we have symmetric A, so

$$R^{-1}A_1C^{-1} = A = A' = C^{-1}A_1'R^{-1}, \tag{2A.4}$$

$$A_1' = C'A'R' = CAR = CR^{-1}A_1C^{-1}R. \tag{2A.5}$$

But both A_1' and A_1 are doubly stochastic, as also is IA_1I, so by the uniqueness part of Sinkhorn's theorem, the diagonal matrices CR^{-1} and $C^{-1}R$ are scalar multiples of I, whence $C = cR$ for some scalar c which we may and shall choose to be 1. Thus, we have the required form CAC for symmetric A.

Proving the existence part of Sinkhorn's theorem in the case of a symmetric $n \times n$ matrix $A = (a_{jk})$ proceeds via a sequence of vectors $\{\beta^{(r)} : r = 0, 1, \ldots\}$ defined recursively by $\beta^{(0)} = \mathbf{e}$ and

$$\beta_j^{(r+1)} = \frac{1}{\sum_{k=1}^n a_{jk}\beta_k^{(r)}}\,. \tag{2A.6}$$

From Seneta's Lemma 2.5 we have for $r = 1, 2, \ldots$

$$\min_j \frac{\beta_j^{(2r)}}{\beta_j^{(2r-2)}} \le \min_j \frac{\beta_j^{(2r-1)}}{\beta_j^{(2r+1)}} \le \min_j \frac{\beta_j^{(2r+2)}}{\beta_j^{(2r)}} \le \min_j \frac{\beta_j^{(2r+1)}}{\beta_j^{(2r+3)}}\,. \tag{2A.7}$$

The matrices $A^{(r)} \equiv \beta^{(r)}A\beta^{(r-1)}$ are specified element-wise as

$$\left(\beta_j^{(r)}a_{jk}\beta_k^{(r-1)}\right) \tag{2A.8}$$

so using the definition at (2A.6) of $\beta^{(r)}$ each of its row sums equals 1, i.e., $A^{(r)}$ is row stochastic. Monotonicity and stochasticity show that a limit matrix and limit vector $\lim_{r\to\infty}\beta^{(r)}$ exist also, and by inspection, the latter is β as required for $B = \mathrm{diag}(\beta)$.

Now apply these results to the matrix $A = \Sigma_X$ with

$$\beta^{(0)} = \mathbf{e}, \quad X_i^{(r)} = B^{(r)}X_i\,, \quad a_{jk} = \mathrm{cov}(X_{ij},\, X_{ik}), \tag{2A.9}$$

so that

$$\begin{aligned}\frac{1}{\beta_j^{(r+1)}} &= \sum_{k=1}^n a_{jk}\beta_k^{(r)} = \sum_{k=1}^n \mathrm{cov}(X_{ij},\, X_{ik})\beta_k^{(r)} = \sum_{k=1}^n \mathrm{cov}(X_{ij},\, X_{ik}^{(r)}) \\ &= n\ \mathrm{cov}(X_{ij},\, \bar{x}_{ik}^{(r)})\end{aligned} \tag{2A.10}$$

where $\bar{x}_i^{(r)}$ is the mean of the scores X_{ij} using the r^{th} approximation to the scaling vector β. Each mean converges to $\bar{y}_i$, and so we recover the equations (2.16) up to the multiplicative constant $\mathrm{var}(\bar{y}_i)$.

Chapter 3

Modelling Course Scores

"[The model of Daley and Seneta] makes explicit the intention of the entire activity: to use scores on different courses to estimate a value v_i on a single underlying variable for each student i." (Masters & Beswick, 1986, §2.46)

True Score Modelling

As already noted concerning the idealized data set $\mathcal{X}$ of Chapter 2, it has long been a common observation that students who achieve higher scores in any one course tend to do so in others, and it matters not whether achievement is measured via school-based assessments or public examinations. In data analytic terms, sets of course scores in most subjects tend to be positively correlated. For over two or more decades this has prompted many Australian workers[7] to describe course scores by a one-factor model. This postulates that the raw scores X_{ij} in any given course j can be transformed into scores Y_{ij} which preserve the rank ordering and can be expressed as

$$Y_{ij} = v_i + e_{ij} \tag{3.1}$$

for some common factor v_i that may be described as "relative general achievement" and error terms e_{ij} that are uncorrelated with both $\{v_i\}$ and error terms for other courses, and have zero mean and standard deviation σ_j^2. These error terms may comprise both model-fit error and measurement error.

The inclusion of an "error" term is a feature of the "true score" model seen most frequently in the literature in connection with the scores obtained in standardized testing (e.g. Chapter 2 of Lord & Novick, 1968). This model for examination marks or mental test scores postulates that, if students $i = 1, \ldots, n$ are subjected to (hypothetical) repeated determinations of their scores X_{ij} in course j by replication of the "run" (i.e., the whole procedure of determining those scores), the resulting set of scores may be represented as

$$\begin{aligned}(X_{ij} \text{ in run } r &\equiv (\text{observed score for } i \text{ in course } j \text{ in run } r) \\ &= (\text{true score for } i \text{ in course } j) + (\text{error in run } r),\end{aligned} \tag{3.2}$$

where the error terms are random variables independent both between runs and of the true score, and have mean zero and variance s_j^2. In practice, we have observations from a single run. What the model does is to provide a useful interpretation for $\{X_{ij}\}$ which, when put in the setting of an aggregate score like T_i at (2.1) or the sum at (2.18), yields respectively

$$T_i = \sum_{j=1}^{n} (\text{true score for } i \text{ in course } j) + \sum_{j=1}^{n} (\text{error in } X_{ij})\,, \tag{3.3}$$

$$\sum (X_{ij} - \xi_j)/\tau_j = \sum_{j \text{ in } \mathcal{S}_i} (\text{true score for } i \text{ in course } j)/\tau_j + (\text{const.}) + (\text{error}) \tag{3.4}$$

[7] For example, Aitkin, 1968; Cook & Cooney, 1976; McGaw, 1977; Daley & Seneta, 1986. There is a short bibliography of work outside Australia in the last reference. Note also Manly (1988).

where $\mathcal{S}_i$ denotes the subset of courses in which i has scores. There is one obvious difference between (3.3) and (3.4) in the inclusion of scaling factors τ_j in the latter. This is of no consequence because the argument of Chapter 2 and the true score model apply equally to replacing the "true score" terms in (3.3) by any transformed scores such as Y_{ij}. Indeed, the argument of Chapter 2 points precisely to the use of such a modified form of (2.1) or (3.3).

One-factor Models and Parameter Estimation

The analyses of the principal component approach of Chapter 2 can now be linked with true score models in a variety of ways. The simplest is to identify the true scores appearing in (3.4), or in (3.3) after scaling, with the common factor v_i of (3.1) and $V_{i1}/\sqrt{n}$ of (2.13), so that the error term e_{ij} of (3.1) is identified with e'_{ij} :

$$e'_{ij} = \sum_{k=2}^{n} \ell_{jk} V_{ik} \,. \tag{3.5}$$

Then

$$\operatorname{var}(e'_{ij}) = \sigma_j^2 = \sum_{k=2}^{n} \ell_{jk}^2 \lambda_k \,, \tag{3.6a}$$

$$\operatorname{cov}(e'_{ir},\, e'_{is}) = \sum_{k=2}^{n} \ell_{rk}\ell_{sk}\lambda_k \quad (r \neq s), \tag{3.6b}$$

whereas the model at (3.1) assumes that $\operatorname{cov}(e_{ir},\, e_{is}) = 0$ $(r \neq s)$. From (3.1) we have

$$\operatorname{var}\left(\textstyle\sum_{j=1}^{n} e_{ij}\right) = \sum_{j=1}^{n} \operatorname{var}(e_{ij}) = \sum_{j=1}^{n} \sigma_j^2 \,, \tag{3.7}$$

while from (3.5) and (3.6) the corresponding variance would be zero because the sum over $j = 1, \ldots, n$ of the right-hand side of (3.5) is zero identically in i. The source of the inconsistency between the principal component representation and the model (3.1) lies in the inclusion in $V_{i1}\sqrt{n}$ of part of what is regarded as error in (3.1).

Using the model (3.1) to describe the scaled scores, and confining attention to linear transformations

$$Y_{ij} = \alpha_j + \beta_j X_{ij} \,, \tag{3.8}$$

how should the parameters $\{(\alpha_j,\, \beta_j)\}$ and $\{v_i\}$ be estimated? There are three approaches detailed specifically in Daley & Seneta (1986), and an allusion to a preferred, fourth method. What should be borne in mind is that the aim of the exercise is to produce an estimator for $\{v_i\}$, and that methods that may be easy or optimal for the rescaling parameters $\{(\alpha_j,\, \beta_j)\}$ may not necessarily be optimal for $\{v_i\}$.

A simple approach to the problem is to try and estimate $\{v_i\}$ independently of the course scores $\{X_{ij}\}$. This is what is done in Western Australia for example where raw scores from the Australian Scholastic Aptitude Test (ASAT) are taken as estimators of $\{v_i\}$. In the context of producing an aggregate, it is an inefficient procedure, and unnecessarily introduces observable imprecision that is extraneous to the use of the achievement scores $\{X_{ij}\}$ to furnish a measure of general achievement. In fairly simple terms, when a student has (say) five course scores and an ASAT score, this approach takes the latter as defining v_i for scaling purposes with either no error or constant error variance, while the former are taken as the scores that matter when determining v_i for implementation. Typically, five scores contain more information than one, so that scaling parameters when determined from ASAT scores alone incorporate much more variability than is necessary. Worse again, there is distortion between the scale used for different averages and the scales that the various standard deviations represent (and these last use mutually inconsistent scales as well!).

Method-of-Moment Estimation

The standard statistical approaches of fitting by Least Squares and Maximum Likelihood were outlined in Daley & Seneta (1986), and were rejected on the grounds of representing projections of the scores on the nominal scale $\{V_{i1}\}$, and requiring assumptions which were so strong as to represent an undue imposition of conformity on the data. (The problem of projections is a more acute version of the "unequal correlation" property underlying the observation in McGaw (1987) noted below (2.12).)

> " ... The pragmatic approach to estimating the quantities $(a_j,\, b_j)$ is to adopt a moment method approach ... From a statistical viewpoint, this moment method approach appears to be the one with the most cogent and assumption-free arguments underlying it." (Daley & Seneta, 1986, p.152)

The parametric specification used at (3.8) differs slightly from that of the original exposition, but is consistent with subsequent discussions.

Starting from equations (3.1) and (3.8), and using a subscript on formal moment operators to denote that the moment is confined to the subset of scores from the course j or individual i concerned, it follows from $v_i = \mathrm{E}(Y_{ij}) = \alpha_j + \beta_j \mathrm{E}(X_{ij})$ that

$$\begin{aligned} \mathrm{ave}_j(v_i) &= \alpha_j + \beta_j \,\mathrm{ave}_j\big(\mathrm{E}(X_{ij})\big), && (3.9)\\ \mathrm{var}_j(v_i) &= \beta_j \,\mathrm{cov}_j\big(v_i,\, \mathrm{E}(X_{ij})\big), && (3.10)\\ v_i &= \mathrm{ave}_i\big(\alpha_j + \beta_j \mathrm{E}(X_{ij})\big), && (3.11)\\ \mathrm{var}_j(v_i) &= \beta_j^2 \,\mathrm{var}_j\big(\mathrm{E}(X_{ij})\big) + \sigma_j^2\,. && (3.12) \end{aligned}$$

The Method-of-Moment estimation procedure consists simply of replacing the expectations $\mathrm{E}(X_{ij})$ in equations (3.9)–(3.11) by the scores X_{ij} themselves, and solving the resulting equations. Inspection of (3.11) shows that it is the same as estimating v_i by the average scaled score Y_{ij}. This means that we would identify such v_i with the average score $\bar{y}_i$ of Chapter 2 if all students were following precisely the same curriculum; then also, equation (3.10) is the same as equation (2.16).

CONCLUSION 3.1. **The Method-of-Moment estimation procedure for finding scale parameters yields the same estimators as under a principal component analysis when the data set is a balanced set as in Chapter 2.**

A general weakness of previous published accounts of scaling procedures has been the absence of a connected discussion of both the location and scale parameters that should be used. After all, if different groups have their scores spread according to the means of certain "scaling criterion" scores, then in spreading scores within a group via their standard deviation care should be taken to ensure that the scales used for the two spreading operations are consistent. (3.9)–(3.11) do not have this weakness.

Two-factor Models

So far we have used a one-factor model. It has been a common observation (e.g. Cooney, 1976; pp.142–146 of Beswick, Schofield, Meek and Masters, 1985) that course scores require for their representation at least a two-factor model. In terms of the principal component representation in the setup of Chapter 2, this means that both $\{V_{i1}\}$ and $\{V_{i2}\}$ can be interpreted meaningfully in terms of relative achievement. What is also commonly found in such studies is that a two-factor model generally suffices to describe common features of the data set: the rest can be left to a mix of "error" and so-called "unique" factors, with rather less required for the latter granted the nature of identifiable measurement error[8] when the scores concerned come from external exams.

We therefore consider what happens to our analysis when, instead of the relation

$$\text{(true score for } i \text{ in course } j) = v_i \tag{3.13}$$

implied by (3.1) and (3.2), we have

$$\text{(true score for } i \text{ in course } j) = v_i + \gamma_j v_{i2} \tag{3.14}$$

for some unknown constants $\{\gamma_j\}$ and second factors $\{v_{i2}\}$ that are uncorrelated with the first factors $\{v_i\}$ and error terms. This matter was studied in Daley (1988) where it was assumed that students tend to choose more courses amongst those where they are relatively stronger. In terms of (3.13) this means choosing courses j for which v_{i2} has the same sign as γ_j, so that students' aggregate scores, formed from the course scores in whatever courses they happen to take, are represented by

$$\begin{aligned}\text{(student } i\text{'s aggregate score)} &= \sum_{j \text{ in } \mathcal{S}_i} (v_i + \gamma_j v_{i2} + \text{error}) \\ &= \text{(no. of courses)} v_i + C_i |v_{i2}| + \text{error} \end{aligned} \tag{3.15}$$

where the coefficients C_i are much smaller than the coefficient of v_i and mostly positive, and the variance $\operatorname{var}(v_i)$ of $\{v_i\}$ over the population is much larger than that of $\{v_{i2}\}$, and larger still than $\operatorname{var}\big([C_i/\text{(no. of courses)}]|v_{i2}|\big)$.

Next, ask what happens in numerically fitting a one-factor model

$$\text{(true score for } i \text{ in course } j) = v_i^{(1)} + e_{ij}^{(1)} \tag{3.16}$$

to a data set that in fact has a two-factor structure as at (3.14). Then

$$\text{(estimate of } v_i^{(1)}) = \text{(no. of courses)} v_i + C_i |v_{i2}|. \tag{3.17}$$

Use characteristic values for the relative variability of $\{v_i\}$ and $\{v_{i2}\}$, and use over-estimates for $\{\gamma_j\}$. Daley (1988) showed that these two modifications to the one-factor model with all students taking all courses gives estimates of scale factors β_j that may be biased up to 1%, which is far smaller than the 5% or more of a two-moment scaling procedure and 10% or so from a least squares procedure.

CONCLUSION 3.2. **For practical purposes, fitting a data set having a two-factor structure as at (3.14) by the Method-of-Moment estimation procedure valid for scaling a one-factor model results in negligible differences in the aggregate scores produced from the original and scaled data sets.**

[8] Daley (1985b) analysed data from NSW *HSC Examination Statistics* and deduced estimates of the errors between examination marks and school-based estimates of those marks. Some of these analyses were updated in Daley & Eyland (1987) when estimates were replaced by school-based assessments.

We can also gauge the practical effects of this model mis-specification from the use on the same set of NSW HSC data of two Other Course Score scaling procedures, the University of Sydney procedure for NSW UCAC, and Method-of-Moment procedure by the Canberra CAE for itself (Daley, 1987b). The differences in the estimates of $\{\beta_j\}$ between these two procedures may be up to 5% to 7% (Daley, 1987a and 1988a). The rankings given to students under one procedure rather than the other differ by rather less than 1 percentile point around the median. Such a discrepancy in classification of students, equivalent to acceptance or rejection of an application for admission to a course, is far smaller than what has hitherto been unknown and accepted in ignorance (cf. also Table 12.4(b)).

While Masters and Beswick recognized the need for any scaling or aggregation procedure to be supervised by a statistical model, and saw that a one-factor model provides a context where the aim of aggregation is plainly given as the estimation of a parameter in such a model, they still failed[9] to recognize in their report that since

> "the main purpose of the model [in this paper] ... is to provide a framework within which different moderation procedures can be compared as statistical procedures." (Daley & Seneta, 1986, p.144),

the model of necessity "supervises" each of the procedures. What the theory of Chapters 2 and 3 does is to use the principles endorsed by Masters and Beswick to demonstrate the nature of an aggregate score and to show that, within the types of data set that arise, the modelling of Daley & Seneta (1986) goes much further than stating the object of the exercise, justified by Chapter 2: its extension in Daley (1988) establishes that the unsubstantiated fears raised by Masters and Beswick under an umbrella of multidimensionality are in fact groundless. In Chapters 6 and 7, we shall see that their partial persuasion of the 1986 Review Committee that wrote *Making Admission to Higher Education Fairer (= MATHEF)* that the sex bias problem giving rise to the Committee's existence was a consequence of multidimensionality, is groundless also: the evidence points to its being a consequence of different educational measurement properties of different groups of individuals on different assessments.

[9] Also, in their §2.49 Masters & Beswick misquote from Daley & Seneta (1986), by attributing comments about least squares estimators as applying to Method-of-Moments estimators.

Chapter 4

Constructing and Using an External Reference Scale

In the simple setup of Chapter 2 all students are assessed in common, and the discussion there applies as well to the internal assessments of any particular school as to the results of a common set of (external) examinations taken over a much larger region. This chapter discusses possible modifications to this simpler setup needed to cope with systems like those of the ACT and Queensland for which the simpler setup applies only to each school that is a member of the much larger system. These two educational systems use both school-based course scores and scores from the Australian Scholastic Aptitude Test (ASAT) to construct a Tertiary Admission Index (TAI) for each student seeking such. The systems claim that these scores reflect school-based measures of general relative achievement fairly for all students. The arguments we give reflect partly on the way the systems presently operate, and partly on general principles that indicate what methodology may best lay claim to the statement that the TAI so constructed is a fair system-wide measure of general relative achievement.

Constructing an Optimum Reference Scale

Suppose that estimates $\{V_{i1}\}$ of measures of relative general achievement have been produced *within* schools of such a system, and that all students have a set of scores $\{(A_{iQ}\ \ A_{iV}\ \ A_{iW})\}$ constructed *system-wide*. How best can we use these scores to place the various sets of scores $\{V_{i1}\}$ on a system-wide scale? There are two parts to this question:

(a) What combination of the scores $\{(A_{iQ}\ A_{iV}\ A_{iW})\}$ best resembles $\{V_{i1}\}$?

(b) How should such a combination be used?

We note here a couple of facts for the record. The standard educational approach to the first part of the problem, as for example in NSW or WA in the context where the scores $\{V_{i1}\}$ are school-based scores in a particular subject or course and whole families of school-based scores like $\{V_{i1}\}$ exist, one family to each course, has been to presume that a common public examination on a syllabus that is sufficiently close to the school-based syllabus will produce scores that are adequate for the purpose. The measure of adequacy commonly reported (but, it is not the most appropriate measure to report) is the correlation coefficient between school- and exam.-based scores. This approach was adopted in Queensland in the sense that there, total ASAT scores $\{A_i\}$ are used, and the ACT initially did likewise. Then *c.* 1984 the ACT sought to control the composition of $\{A_i\}$ psychometrically, replacing these raw scores by a 50 : 50 mixture of Quantitative and Verbal sub-scale scores $\{(A_{iQ}\ A_{iV})\}$. Since the sub-scale scores are constructed by statistical devices, it seems only proper and consistent to use similar devices, at least initially, to answer part (a) of the present problem concerning the scores $\{(A_{iQ}\ A_{iV}\ A_{iW})\}$ in relation to thirteen sets of scores $\{V_{i1}\}$, one set per college. As a check, the same question can be asked separately of the groups of female and male students within each of the eight mixed-sex colleges.

Desiderata of any reference scale, to be consistent with the principles P 2–4 and the analysis of Chapter 2, should include maximum agreement with the scales $\{V_{i1}\}$ produced within each school. In the present context this means that the correlation between $\{V_{i1}\}$ and whatever is used as $\{A_i\}$ should be maximized; bias questions are deferred to Chapter 7. When a system-wide set of scores like $\{(A_{iQ}\ A_{iV}\ A_{iW})\}$ is available for the construction of such $\{A_i\}$, we ask: within each school,

(i) What is the best "mixture" of these three components when used to predict scores $\{V_{i1}\}$ (and we interpret "mixture" as a linear combination)?

(ii) What is the best global mixture closest to each of the mixtures determined in (i)?

(iii) Does the global mixture produce significantly worse predictions than any of the separate optimal mixtures of (i)?

(iv) When used as a predictor of $\{V_{i1}\}$, does the global mixture have approximately constant unexplained variability (i.e., "noise") across different schools?

In preparing the next four tables, we have followed existing ACT practice in scaling course scores except to replace the ASAT scaling procedure by the Other Course Score procedure of Chapter 3 that agrees with the approach of Chapter 2 in that uniform setup. This yields sets $\{v_i\} \equiv \{V_{i1}\}$, as well as quasi-TE scores which we denote $\{TM_i\}$ (M for Method-of-Moments). We have then done various regression analyses to "predict" either $\{v_i\}$ or $\{TM_i\}$ by linear functions of the triplets $\{(A_{iQ}\ A_{iV}\ A_{iW})\}$, and used standard statistical measures to compare these predictors. For each set of scores within a given college, say $\{v_i\}$, it is feasible to compare different predictors by their correlations, but it is not logical to make comparisons across colleges via correlations because the latter can change considerably with the spread of $\{v_i\}$. The F ratios in the column QVW :*Opt.* compares the simple average of the thirteen college-wise optimum mixtures with the optimum mixture within each college both as a whole and, for each mixed-sex college, for each gender group.

Table 4.1 shows the optimum percentage mix of the component scores (either two or three) used to predict $\{v_i\}$ and $\{TM_i\}$, within each college for all students, and when the college is mixed-sex, for males and females separately. In Table 4.2, each F ratio entry in the column QVW:*Opt.* compares the improved fit that holds when, for the school or gender group within the school concerned, the best linear combination of the three scores is used in place of the average of the college-wise optimal mixtures. Of the 30 regressions shown, 2 have F ratios exceeding the 5% significance level (though only just: 4.7% and 4.2%), so there is no evidence in these analyses indicating that the overall mixture is other than common to all sets $\{V_{i1}\}$ and their gender-based subsets. Thus, in the space spanned by $\{(A_{iQ}\ A_{iV}\ A_{iW})\}$, no set of scores $\{V_{i1}\}$ is concentrated in a direction significantly different from that of the predictor $0.472A_{iQ} + 0.181A_{iV} + 0.347A_{iW}$, i.e., there are no significant differences between colleges in the psychometric composition of aggregates over the college as a whole, whether single- or mixed-sex or within the gender-based groupings in mixed-sex colleges, except possibly in directions orthogonal to the space $\{(A_{iQ}\ A_{iV}\ A_{iW})\}$.

Table 4.2 lists certain results about the different regression analyses. The F ratios comparing predictors on two rather than three components shows that of the three scores concerned, the least useful is the multiple choice Verbal sub-scale score. Even the sub-optimal mixture used in 1986, denoted AST in the Table, was a distinct improvement[10] on the optimal Q–V mixture.

Conclusion 4.1. **The Writing Task used in 1986 was a successful innovation in the sense of enabling the construction of a better reference scale.**

The detail of the various mixtures given in Table 4.1 indicates the scope for variation about an optimum predictor that may be observed in practice. What may seem surprising is that, in moving from the three-component predictor to the two-component predictor based on $\{(A_{iQ}\ A_{iW})\}$, the explanatory power of the Verbal sub-scale score $\{A_{iV}\}$ is taken up rather more by $\{A_{iQ}\}$ than by Writing Task scores $\{A_{iW}\}$. This result is consistent with the existence of an observable mode-of-assessment effect on the determination of "Verbal" ability, as conveyed by an essay-writing exercise like the Writing Task as opposed to the multiple-choice methods used to find A_{iV}.

[10] The standard cautionary remark about a conclusion being based on just one ASAT paper and Writing Task should be added here.

Table 4.1

Mixture ratios of optimal predictors of v_i and TM_i.

Run #	Regressing $\{v_i\}$			Regressing $\{TM_i\}$
	Q : V : W	Q : W	Q : V	Q : V : W
		(a) All Students		
620	62.5 : 17.9 : 19.6	77.6 : 22.4	77.4 : 22.6	60.9 : 19.4 : 19.7
621	51.4 : 15.5 : 33.2	61.7 : 38.3	67.8 : 32.2	51.2 : 14.2 : 34.6
622	57.5 : 19.4 : 23.1	72.6 : 27.4	71.2 : 28.8	60.1 : 18.1 : 21.8
623	51.0 : 8.4 : 40.6	56.8 : 43.2	70.0 : 30.0	50.7 : 10.7 : 38.5
624	20.6 : 42.6 : 36.8	52.3 : 47.7	35.6 : 64.4	23.7 : 41.0 : 35.3
625	46.7 : 4.6 : 48.6	49.6 : 50.4	71.9 : 28.1	43.4 : 7.7 : 48.8
626	41.3 : 16.9 : 41.8	52.2 : 47.8	58.5 : 41.5	42.1 : 16.7 : 41.1
627	57.7 : 16.1 : 26.2	71.1 : 28.9	76.2 : 23.8	59.9 : 15.7 : 24.4
628	36.4 : 20.6 : 43.0	53.3 : 46.7	72.0 : 28.0	36.7 : 20.4 : 42.9
629	54.0 : 10.1 : 35.9	61.1 : 38.9	66.8 : 33.2	55.6 : 10.3 : 34.2
630	50.7 : 8.2 : 41.1	55.8 : 44.2	57.3 : 42.7	51.7 : 7.5 : 40.8
631	44.3 : 28.7 : 27.0	64.5 : 35.5	54.2 : 45.8	44.3 : 27.6 : 28.1
632	39.2 : 26.5 : 34.3	56.6 : 43.4	51.0 : 49.0	41.6 : 29.3 : 29.1
		(b) Female Students Only		
621	76.7 : –8.1 : 31.4	71.2 : 28.8	98.6 : 1.4	69.3 : –5.6 : 36.2
622	66.3 : 10.9 : 22.8	75.5 : 24.5	82.8 : 17.2	65.4 : 12.2 : 22.4
623	58.5 : 3.6 : 37.9	61.1 : 38.9	83.2 : 16.8	56.1 : 6.4 : 37.5
625	41.1 : 26.5 : 32.4	57.1 : 42.9	50.0 : 50.0	39.7 : 29.2 : 31.1
626	39.5 : 14.5 : 46.0	49.5 : 50.5	66.3 : 33.7	35.6 : 16.9 : 47.4
629	52.3 : 5.7 : 42.0	56.4 : 43.6	69.2 : 30.8	50.7 : 6.2 : 43.2
630	46.7 : 3.7 : 49.7	48.8 : 51.2	63.8 : 36.2	48.3 : –1.1 : 52.8
632	39.4 : 31.8 : 28.8	61.9 : 38.1	51.8 : 48.2	40.4 : 35.3 : 24.3
620	62.5 : 17.9 : 19.6	77.6 : 22.4	77.4 : 22.6	60.9 : 19.4 : 19.7
628	36.4 : 20.6 : 43.0	53.3 : 46.7	72.0 : 28.0	36.7 : 20.4 : 42.9
631	44.3 : 28.7 : 27.0	64.5 : 35.5	54.2 : 45.8	44.3 : 27.6 : 28.1
		(c) Male Students Only		
621	48.5 : 20.6 : 30.9	64.6 : 35.4	60.9 : 39.1	51.6 : 16.9 : 31.5
622	60.9 : 21.9 : 17.2	79.6 : 20.4	73.5 : 26.5	63.3 : 20.5 : 16.3
623	50.5 : 11.4 : 38.1	58.7 : 41.3	69.8 : 30.2	50.0 : 14.3 : 35.7
625	91.2 : –24.2 : 33.0	72.5 : 27.5	119.9 : –19.9	87.0 : –22.2 : 35.2
626	48.8 : 18.5 : 32.7	61.5 : 38.5	67.3 : 32.7	55.0 : 15.6 : 29.4
629	64.9 : 7.8 : 27.3	70.7 : 29.3	78.2 : 21.8	66.9 : 9.1 : 24.0
630	61.4 : 5.7 : 32.9	65.4 : 34.6	74.3 : 25.7	62.6 : 7.7 : 29.8
632	63.4 : 11.9 : 24.7	71.8 : 28.2	76.2 : 23.8	67.7 : 13.6 : 18.7
624	20.6 : 42.6 : 36.8	52.3 : 47.7	35.6 : 64.4	23.7 : 41.0 : 35.3
627	57.7 : 16.1 : 26.2	71.1 : 28.9	76.2 : 23.8	59.9 : 15.7 : 24.4
		(Unweighted) Average Mixture		
(a)	47.2 : 18.1 : 34.7	60.4 : 39.6	63.8 : 36.2	47.8 : 18.4 : 33.8
(b)	51.2 : 14.2 : 34.6	61.5 : 38.5	69.9 : 30.1	49.8 : 15.2 : 35.0
(c)	56.8 : 13.2 : 30.0	66.8 : 33.2	73.2 : 26.8	58.8 : 13.2 : 28.0

TABLE 4.2

Residual (error) mean squares and F ratios

Run #	df	Regressing v_i		Regressing TM_i		F Ratios			
		QVW	Unif.Opt	QVW	AST	QVW:Opt	QVW:QW	QVW:QV	TM_i
				(a) All Students					
620	111	191.9	194.3	2716.	2737.	1.702	2.750	3.130	1.436
621	176	179.5	178.0	2614.	2723.	0.256	3.351	18.148	4.700
622	180	208.5	211.7	3031.	3091.	2.374	5.731	10.302	2.801
623	244	157.4	157.9	2302.	2548.	1.332	1.717	61.674	14.145
624	68	216.6	220.1	3010.	3074.	1.555	8.060	8.575	1.746
625	83	150.8	150.4	2024.	2203.	0.876	0.117	15.311	4.756
626	264	226.6	226.2	3494.	3720.	0.798	5.928	48.165	9.612
627	117	182.4	183.4	2666.	2759.	1.329	3.102	12.034	3.070
628	58	171.8	168.7	2391.	2813.	0.466	4.554	23.815	6.300
629	235	239.7	239.0	3282.	3495.	0.639	1.563	37.081	8.694
630	202	163.2	162.7	2475.	2644.	0.661	0.808	34.997	7.975
631	118	188.1	188.7	2875.	2907.	1.198	9.954	14.657	1.669
632	128	185.4	183.4	2679.	2682.	0.299	5.486	13.130	1.064
				(b) Female Students Only					
621	94	150.3	157.0	2352.	2537.	3.150	0.333	6.200	4.775
622	76	159.0	159.1	2411.	2424.	1.015	0.461	2.612	1.210
623	130	155.8	157.3	2154.	2441.	1.619	0.174	28.993	9.791
625	40	115.0	110.2	1550.	1518.	0.122	2.597	3.718	0.558
626	138	203.6	203.6	3075.	3363.	0.980	2.497	27.787	7.560
629	122	208.4	207.4	2724.	3055.	0.693	0.229	27.376	8.528
630	107	131.1	131.2	1984.	2190.	1.054	0.068	19.357	6.659
632	70	143.1	140.6	2146.	2097.	0.364	3.451	3.706	0.187
620	111	191.9	194.3	2716.	2737.	1.702	2.750	3.130	1.436
628	58	171.8	168.7	2391.	2813.	0.466	4.554	23.815	6.300
631	118	188.1	188.7	2875.	2907.	1.198	9.954	14.657	1.669
				(c) Male Students Only					
621	78	199.6	195.5	2809.	2904.	0.191	3.911	10.645	2.353
622	100	226.7	234.7	3266.	3309.	2.796	5.385	3.849	1.676
623	110	153.5	151.4	2450.	2572.	0.247	1.244	20.642	3.780
625	39	142.9	160.0	1814.	2321.	3.453	1.953	3.927	6.734
626	122	255.9	252.1	3986.	4056.	0.061	3.065	11.126	2.090
629	109	270.9	274.3	3799.	3979.	1.685	0.479	9.991	3.623
630	91	192.9	192.8	2917.	3032.	0.974	0.222	11.861	2.828
632	54	219.1	217.6	3048.	3087.	0.808	0.665	4.203	1.360
624	68	216.6	220.1	3010.	3074.	1.555	8.060	8.575	1.746
627	117	182.4	183.4	2666.	2759.	1.329	3.102	12.034	3.070

The regression error mean squares shown in Table 4.2 are given in the scale of the predicted variables, rather than the predicting variables. Mean squares in the latter scale are given in Table 4.4, the intention being that vertical comparisons within a column should be possible there. Within each of the three groupings, the range of these mean squares is about the same for each of the four regressions, being about 1.5 for Male Students only, 1.8 for Female Students only, and about 1.7

Table 4.3

Correlation Coefficients of General Achievement measures and Linear functions of ASAT sub-scale measures

Run #	Regressing v_i				Regressing TM_i	
	QVW	QW	QV	Unif.Opt	QVW	ASAT
			(a) All Students			
620	0.704	0.695	0.694	0.693	0.696	0.686
621	0.654	0.646	0.607	0.653	0.647	0.623
622	0.701	0.690	0.680	0.692	0.700	0.689
623	0.689	0.686	0.584	0.685	0.693	0.648
624	0.692	0.646	0.643	0.675	0.692	0.672
625	0.614	0.613	0.511	0.603	0.621	0.562
626	0.690	0.682	0.617	0.688	0.674	0.644
627	0.737	0.729	0.705	0.730	0.741	0.725
628	0.797	0.779	0.697	0.794	0.807	0.759
629	0.705	0.702	0.646	0.703	0.733	0.710
630	0.637	0.635	0.550	0.634	0.634	0.595
631	0.776	0.754	0.743	0.770	0.764	0.757
632	0.629	0.608	0.577	0.627	0.623	0.615
			(b) Female Students Only			
621	0.674	0.672	0.646	0.646	0.640	0.592
622	0.615	0.612	0.597	0.601	0.612	0.595
623	0.724	0.723	0.646	0.715	0.739	0.691
625	0.692	0.667	0.655	0.689	0.706	0.696
626	0.705	0.699	0.629	0.700	0.697	0.655
629	0.741	0.741	0.670	0.738	0.770	0.732
630	0.605	0.605	0.502	0.595	0.611	0.543
632	0.540	0.507	0.505	0.534	0.532	0.528
620	0.704	0.695	0.694	0.693	0.696	0.686
628	0.797	0.779	0.697	0.794	0.807	0.759
631	0.776	0.754	0.743	0.770	0.764	0.757
			(c) Male Students Only			
621	0.683	0.663	0.627	0.681	0.684	0.660
622	0.760	0.745	0.749	0.744	0.756	0.746
623	0.648	0.643	0.558	0.646	0.635	0.602
625	0.691	0.672	0.652	0.621	0.702	0.564
626	0.682	0.672	0.645	0.681	0.662	0.648
629	0.681	0.679	0.644	0.668	0.717	0.694
630	0.688	0.687	0.636	0.680	0.683	0.658
632	0.727	0.723	0.701	0.717	0.729	0.712
624	0.692	0.646	0.643	0.675	0.692	0.672
627	0.737	0.729	0.705	0.730	0.741	0.725

for All Students. The mean squares themselves can be tested for homogeneity. In the case of the regressions on the triplet, these lead to crude Bartlett test statistics of 51.9 on 12 df for all students, 38.4 on 10 df for female students, 20.8 on 9 df for male students, and 14.5 on 5 df for non-Government colleges. On the assumption of normality for the errors, these test statistics indicate inhomogeneity strongly for the first two and weakly for the latter two (*c.* 1% significance level). The last statistic is consistent with a long-term observation that correlations between ASAT and TE scores tend to be lower inside the Government secondary colleges than outside. The weighted residual mean squares in Table 4.4 indicate, since $\mathrm{var}(A_i) \approx \mathrm{var}(1.28\times$ [Opt. mixture]$) = 625 = 25^2$, is that 40% to 50% of the variability of even the optimal mixture of a Total ASAT score is unexplained variability with respect to relative general academic achievement. This range corresponds to a signal to noise ratio of about 1.5 to 1.0, whereas the smallest such ratio observed in estimating $\{v_i\}$ in the scaling process is about 1.6, and for several colleges exceeds 3.0. This reflects the greater coherence of school-based course scores amongst themselves as opposed to that of any combination of ASAT scores with even the aggregate of courses scores.

The implication of this analysis is that the choice of the particular method of using any set of Reference Scale scores affects the relative placement of colleges, to the relative advantage of some students (and therefore disadvantage of some others) simply on account of the college they happen to have attended. As will emerge from other analyses (Chapters 8 and 10), an Other Course Score scaling procedure certainly reduces these residual mean squares, while from Table 4.2 it follows that an optimally weighted mixture does rather better than the 1986 Total ASAT scores.

Correlation coefficients of ASAT and TE scores are given in Table 4.3 for those who prefer to compare regression data this way. Such coefficients are less informative, and are harder to use in secondary analyses. In particular, the only meaningful comparisons that can be made without further data are *across* any row, and then only for the same regressing variable; vertical comparisons (i.e., between colleges) are not valid. Nevertheless, the higher correlations consistently reported for single-sex colleges in *Year 12 Study* tables, are not mere artefacts coming from different ranges of scores: they do reflect a closer coherence between ASAT and course scores there.

Since a set of TE scores constitutes a single ranking of student achievement, it is certainly inconsistent with any model of course scores to use more than one reference scale as in the 1986 ACT procedure where, depending on the course, one of ASAT$-Q$, $-V$, and $-T$ is used. This illogicality also introduces selection bias effects (see Chapter 12). Further, the correlations of course scores wuth Other Course Scores are almost invariably rather higher than with *any* combination of ASAT scores.

Conclusion 4.2. **Each year, the optimal predictive combination of system-wide reference scores should be constructed to produce a single system-wide reference measure of general ability in the "dimension" of school-based general achievement measures.**

Note that in 1986, the optimal mixture of the scores $\{(A_{iQ}\ A_{iV}\ A_{iW})\}$ was a 47 : 18 : 35 mixture.

Conclusion 4.3. **The coherence between course scores and even optimally constructed "Total ASAT score" to be used as a reference scale, varies significantly between colleges.**

Using Reference Scale Scores

We now turn to the rather thornier problem of how to use these optimally constructed reference scores when they are regarded as measures of developed academic ability and are assumed to correlate "well" with the measures $\{V_{i1}\}$ within each school (educational measurement workers would look for an "equating" procedure answer to this problem). If indeed the scores correlate perfectly, then there is no concern that it is correct to rescale each set $\{V_{i1}\}$ via parameters (a, b) say such that the first two moments of $\{a + bV_{i1}\}$ and $\{A_i\}$ coincide for each set. In practice we are confronted with somewhat smaller correlations, typically in the range 0.6 to 0.75, with the standard deviation of $\{A_i\}$ for each school similar to its value in the whole population (meaning, the standard deviations are usually within 10%, and certainly 20%, of this global figure).

Table 4.4

Residual Mean Squares in ASAT or Unif. Opt. Scale of General Achievement Measures Predicted from ASAT Sub-scale Scores

Run #	df	Regressing v_i		Regressing TM_i	
		Q–V–W	Unif.Opt.	*Q–V–W*	ASAT
		(a) All Students			
620	111	292.2	286.0	298.7	306.7
621	176	324.5	311.7	329.6	346.9
622	180	321.8	315.3	322.7	332.3
623	244	286.3	293.3	283.3	316.2
624	68	344.4	368.1	344.4	362.4
625	83	331.6	338.9	327.0	364.2
626	264	387.8	388.3	404.0	433.3
627	117	257.1	256.4	253.8	267.0
628	58	214.9	238.1	205.5	249.8
629	235	387.5	402.6	356.5	382.1
630	202	326.8	330.2	328.9	355.3
631	118	227.0	229.1	237.6	243.7
632	128	358.7	361.1	363.1	369.0
		(b) Female Students Only			
621	94	283.1	316.7	306.2	336.9
622	76	389.0	362.7	391.3	404.1
623	130	281.7	287.1	268.7	309.3
625	40	277.5	247.2	267.1	274.5
626	138	387.8	388.0	396.4	440.2
629	122	389.8	412.7	351.9	401.3
630	107	336.7	355.3	332.8	374.5
632	70	318.4	310.6	322.2	324.1
620	111	292.2	286.0	298.7	306.7
628	58	214.9	238.1	205.5	249.8
631	118	227.0	229.1	237.6	243.7
		(c) Male Students Only			
621	78	292.7	258.1	292.0	309.7
622	100	263.3	280.0	267.1	276.4
623	110	286.9	300.1	295.2	315.4
625	39	259.3	353.5	251.7	338.4
626	122	369.0	382.1	387.5	400.1
629	109	360.5	376.5	326.7	348.5
630	91	295.8	300.8	299.7	318.5
632	54	325.7	358.6	323.7	340.7
624	68	344.4	368.1	344.4	362.4
627	117	257.1	256.4	253.8	267.0
Wtd.m.s. (a)	1984		327.1		
Wtd.m.s. (b)	1064		320.3		
Wtd.m.s. (c)	888		320.2		
do. Non-Govt.	652		281.8		

True Score Modelling

Regarded as observed scores in a true score model, it is easily shown that the estimates of V_{i1}, being aggregate scores, have considerably smaller errors than the ASAT scores A_i. However, we shall see in Chapter 6 that there is a difference in structure between ASAT scores and the school-based measures like V_{i1}. This is both a potential and real source of difficulties for their use, whether in the existing scaling procedure or any other.

Let there be given ASAT scores $\{A_i\}$ and school-based measures $\{V_{i1}\}$. To be consistent with Assumption A 5 about ASAT scores we should seek parameters $(a,\, b)$ such that

$$a + bV_{i1} = v_i + e_{iV}\,, \tag{4.1a}$$

$$A_i = v_i + e_{iA} \tag{4.1b}$$

for "true scores" $\{v_i\}$ of general achievement *cum* ability and zero mean error variables $\{e_{iV}\}$, $\{e_{iA}\}$ with variances s_V^2, s_A^2 respectively. Since

$$t(a + bV_{i1}) + (1-t)A_i \tag{4.2}$$

is an estimator of v_i with error variance $t^2 s_V^2 + (1-t)^2 s_A^2$, we obtain a minimum variance estimator by setting $t = 1/[1 + (s_V/s_A)^2]$. Now $s_V^2 \approx s_A^2/5$ so we may take $t = 0.83$. Equating means and variances via the model at (4.1), we have

$$a + b\,\mathrm{ave}(V_{i1}) = \mathrm{ave}(v_i) = \mathrm{ave}(A_i), \tag{4.3a}$$

$$\begin{aligned} b^2\,\mathrm{var}(V_{i1}) - s_V^2 &= \mathrm{var}(v_i) = \mathrm{var}(A_i) - s_A^2 \\ &= b\,\mathrm{cov}(V_{i1},\, v_i) = \mathrm{cov}(A_i,\, v_i). \end{aligned} \tag{4.3b}$$

The last of these relations, yielding the estimator $b = \mathrm{var}(v_i)/\,\mathrm{cov}(V_{i1},\, v_i)$, is of the same form as equation (2.16) satisfied by elements of the diagonal matrix $B = D_0$. Substituting the estimator with $t = 0.83$ from (4.2), when b at (4.2) equals 1, yields

$$\text{estimate of } b \approx \frac{\mathrm{var}(v_i) + 5s_V^2/6}{\mathrm{var}(v_i) + 5s_V^2/6} = 1, \tag{4.4}$$

whereas the relation $b^2\,\mathrm{var}(V_{i1}) = \mathrm{var}(A_i)$, which is close to what is used in Queensland's "second stage of the scaling procedure" (cf. e.g. McGaw, 1977; Keeves, McBryde & Bennett, 1977), yields

$$\text{estimate of } b \approx \sqrt{\frac{\mathrm{var}(v_i) + s_A^2}{\mathrm{var}(v_i) + s_V^2/6}} \approx \sqrt{\frac{1 + s_A^2/\,\mathrm{var}(v_i)}{1 + s_A^2/[6\,\mathrm{var}(v_i)]}}\,. \tag{4.5}$$

With correlation coefficients between $\{A_i\}$ and $\{V_{i1}\}$ around 0.7 (see Table 4.3), the ratio $s_A^2/\,\mathrm{var}(v_i) \approx 1.0$, so the estimate of $b \approx \sqrt{12/7} \approx 1.3$. In terms of the model, the spread of TAI's introduced by the second stage of scaling muddies considerably the inter-school comparisons of the estimates of relative general achievement produced within the schools. The size of this muddying is sufficient to be noticeable, so from a technical viewpoint, it should be brought more under control if feasible (which it is). We do not discuss whether this muddying is to be regarded as socially desirable either in masking levels of academic achievement as determined by the data produced, or else in biasing such levels depending on the group (school) where the TAI's have been determined. Whichever is the case, it is not in accordance with the principle enunciated at the outset of Chapter 2. We conclude that, irrespective of the reasoning used, the ACT system was correct in 1977 to opt not to use this second stage.

One-factor Model

Given both this inconsistency in estimating b by equating variances as below (4.4) and the fact that it involves a multivariate approach to the data set in constructing $\{V_{i1}\}$ but a bivariate approach in using $\{A_i\}$, consider again equations (4.1). Note in particular that (4.1b) is of the same form as equations (2.13) and (3.1) for the scores Y_{ij} used to construct T_i which estimates a multiple of v_i. In that one-factor model, the error variance terms $\operatorname{var}(e_{ij})$ and $\operatorname{var}(e_{iA})$, are not necessarily conceptually distinguishable, comprising as they do both model-fit and measurement error components. We may therefore consider each ASAT score A_i as a potential estimator of v_i in the same way as any scaled course score Y_{ij} is an estimator of v_i. There is of course a superficial difference: neither the ACT nor Queensland systems uses the ASAT score as a possible contributor to the TAI, but that is no reason for not considering its potential as a contributor to the relative general achievement/ability measure V_{i1} when so far as the scaling procedure is concerned a most desirable aim is to secure as a scaling criterion variable a quantity that has good correlational properties with the data. To this end, we find in practice that a set of course scores $\{X_{ij}\}$ can correlate marginally better or worse with the criterion "sum of scaled scores in all other courses" when ASAT scores are included as one of these other courses.

Interlude: Principal Component Representations and Reference Scale

Consider a set of scores much as in Chapter 2 but where now the scores are the $(n+1)$-vectors $(Y_{i1} \ \cdots \ Y_{in} \ A_i)'$. We take the assumption A 4 that general ability can be validly measured by a mix of content from the humanities, ..., and the sciences, to be equivalent to asserting that when the n courses represent a balanced curriculum, the ASAT scores estimate V_{i1} first and foremost, with negligible weight on a second factor V_{i2} (usually, a quantitative/verbal contrast factor), so that, hopefully, only measurement error in A_i relative to V_{i1} remains. Around (3.14) we noted that a two-dimensional space generally suffices to describe the courses scores, so this last statement is equivalent to asserting that the scores $\{A_i\}$ should be indistinguishable from the other component scores $\{V_{ik} : k = 3, \ldots, n\}$ before introduction of the scores $\{A_i\}$. Analyses in Chapter 6 do not support this sanguine outlook.

Assume here without loss of generality that $\sum A_i = 0 = \sum X_{ij}$ (cf. below equation (2.1)). If these scores are regarded as determining the scale across all schools, then in determining the matrix of multipliers $B = \operatorname{diag}(\{\beta_1, \ldots, \beta_n, \beta_A\})$ we should now use $\beta_A = 1$ in place of $\det(B) = 1$. Equations (2.16) are still to be satisfied where now, after rescaling, we should have

$$\bar{y}_i = \frac{\beta_1 X_{i1} + \cdots + \beta_n X_{in} + A_i}{n+1}, \tag{4.6a}$$

$$\beta_j = \frac{\operatorname{var}(\bar{y}_i)}{\operatorname{cov}(X_{ij}, \bar{y}_i)}. \tag{4.6b}$$

Example 4.1. To gain some insight as to what may happen with extreme versions of scores $\{A_i\}$, consider the case where $\{A_i\}$ has correlation 1.0 with $\{V_{i1}\}$, which implies that it is uncorrelated with all other components $\{V_{ik} : k = 2, \ldots, n\}$, and that $A_i = (\gamma/\sqrt{n})V_{i1}$ for some $\gamma > 0$. From the earlier discussion we would expect that the transformation $Y_i^* = B^*(Y_i' \ A_i)'$ with $B^* = \operatorname{diag}((\gamma \mathbf{e}' \ 1))$ should yield a representation $Y_i^* = L^* V_i^*$ with $\ell_{j1}^* = 1/\sqrt{n+1}$. To check this, observe that

$$B^*(Y_i' \ A_i)' = (\gamma Y_i' \ (\gamma/\sqrt{n})V_{i1})' \tag{4.7}$$

so the covariance matrix Σ^* of $\{Y_i^*\}$ equals

$$\gamma^2 \begin{pmatrix} \Sigma & \operatorname{cov}(V_{i1},\, Y_i)/\sqrt{n} \\ \operatorname{cov}(V_{i1},\, Y_i')/\sqrt{n} & \operatorname{var}(V_{i1})/n \end{pmatrix} = \gamma^2 \begin{pmatrix} \Sigma & (\lambda_1/n)\mathbf{e} \\ (\lambda_1/n)\mathbf{e}' & \lambda_1/n \end{pmatrix} . \tag{4.8}$$

The sum of elements in the j^{th} row of Σ equals

$$\sum_{k=1}^{n} \operatorname{cov}(Y_{ij},\, Y_{ik}) = \operatorname{cov}(Y_{ij},\, (\sqrt{n}\,)V_{i1}) = \lambda_1\,, \tag{4.9}$$

which is independent of j, and $n(\lambda_1/n) = \lambda_1$ so the last column of Σ^* is linearly dependent on the first n columns. Thus the smallest eigenvalue of Σ^* is $\lambda_{n+1} = 0$. It can now be verified that $\Sigma^* = L^*\text{diag}(\{\lambda_k^*\})(L^*)'$ where

$$L^* = \begin{pmatrix} \mathbf{e}/\sqrt{n+1} & \ell_2 & \cdots & \ell_n & \mathbf{e}/\sqrt{n(n+1)} \\ 1/\sqrt{n+1} & 0 & \cdots & 0 & -\sqrt{n/(n+1)} \end{pmatrix} , \tag{4.10a}$$

$$\lambda_k^* = \begin{cases} \gamma^2(n+1)\lambda_1/n & k=1, \\ \gamma^2\lambda_k & k = 2,\dots,n, \\ 0 & k = n+1. \end{cases} \tag{4.10b}$$

Further, if we make the identifications

$$V_{ik}^* = \begin{cases} \gamma\sqrt{(n+1)/n}\,V_{i1} & k=1, \\ \gamma V_{ik} & k = 2,\dots,n, \\ 0 & k = n+1, \end{cases} \tag{4.11}$$

then we have $Y_i^* = L^*V_i^*$. Note that

$$\operatorname{corr}(A_i,\, V_{i1}) = 1 \qquad \text{implies} \qquad \operatorname{cov}(Y_{ik},\, A_i) = \text{const.} \quad (\text{all } k). \tag{4.12}$$

EXAMPLE 4.2. For general $\{A_i\}$ define coefficients m_k $(k = 1,\dots,n)$ by

$$m_k = \operatorname{cov}(A_i,\, V_{ik})/\lambda_k\,. \tag{4.13a}$$

When the right-hand side below does not vanish identically, there is some non-zero m_{n+1} which when fixed defines $\{V_{i,n+1}\}$ by

$$m_{n+1}V_{i,n+1} = A_i - \sum_{k=1}^{n} m_k V_{ik}\,. \tag{4.13b}$$

It is easily checked that any set $\{V_{i,n+1}\}$ defined this way is orthogonal to $\{V_{ik}\}$ for $k = 1,\dots,n$. Equation (4.13b) is equivalent to the representation

$$A_i = \sum_{k=1}^{n+1} m_k V_{ik}\,. \tag{4.13c}$$

Now suppose that $\{A_i\}$ satisfies the latter equality at (4.12), i.e., that

$$\operatorname{cov}(Y_{ik},\, A_i) = C \quad (k = 1,\dots,n) \tag{4.14}$$

for some positive constant C. Then

$$C = \sum_{k=1}^{n+1} m_k \ell_{rk} \lambda_k = \sum_{k=1}^{n} m_k \ell_{rk} \lambda_k \tag{4.15}$$

since in

$$Y_{ir} = \sum_{k=1}^{n+1} \ell_{rk} V_{ik} = \sum_{k=1}^{n} \ell_{rk} V_{ik} \tag{4.16}$$

we have $\ell_{r,n+1} = 0$. Sum over $r = 1, \ldots, n$ and recall that the elements of ℓ_k have sum $\ell_{\cdot k}$ which equals $\sqrt{n}$ for $k = 1$ and vanishes otherwise. Then we have

$$nC = \sum_{r=1}^{n} \sum_{k=1}^{n} m_k \ell_{rk} \lambda_k = \sum_{k=1}^{n} m_k \ell_{\cdot k} \lambda_k = (\sqrt{n}\,) m_1 \lambda_1 \,, \tag{4.17}$$

i.e., $m_1 \lambda_1 = C\sqrt{n}$. Substitute this result in (4.15) to give

$$C = m_1 \ell_{r1} \lambda_1 + \sum_{k=2}^{n} m_k \ell_{rk} \lambda_k = C + \sum_{k=2}^{n} m_k \ell_{rk} \lambda_k \,.$$

Thus, defining $\mu_1 = 0$, $\mu_k = m_k \lambda_k$ $(k = 2, \ldots, n)$, we have $L\mu = 0$. Consequently, since L is orthogonal, $\mu = 0$, i.e., $m_k \lambda_k = 0$ for $k = 2, \ldots, n$, and therefore $\operatorname{cov}(A_i,\, V_{ik}) = 0$ for $k = 2, \ldots, n$.

The covariance matrix of $\{(Y_i'\ A_i)'\}$ is

$$\begin{pmatrix} \Sigma & C\mathbf{e} \\ C\mathbf{e}' & \operatorname{var}(A_i) \end{pmatrix} . \tag{4.18}$$

Consider now the set $\{(\gamma Y_i'\ A_i)'\}$ for some non-zero γ. This has covariance matrix

$$\begin{pmatrix} \gamma^2 \Sigma & \gamma C\mathbf{e} \\ \gamma C\mathbf{e}' & \operatorname{var}(A_i) \end{pmatrix} . \tag{4.19}$$

For the last column to be linearly dependent on the other n columns there must exist a non-zero n-vector μ such that

$$\gamma^2 \Sigma \mu = \gamma C \mathbf{e} \quad \text{and} \quad \gamma C \mathbf{e}' \mu = \operatorname{var}(A_i) \,. \tag{4.20}$$

Each column of Σ sums to λ_1 so the former of these relations implies that, if such μ exists, then $\gamma \lambda_1 \mathbf{e}' \mu = nC$, which with the latter implies that

$$\lambda_1 \operatorname{var}(A_i) = C^2, \quad \text{i.e.,} \quad \operatorname{var}(V_{i1}) \operatorname{var}(A_i) = [\operatorname{cov}(A_i,\, V_{i1})]^2, \quad \text{i.e.,} \quad \operatorname{corr}(A_i,\, V_{i1}) = 1.$$

Taken together, Examples 4.1 and 4.2 show that, unless A_i is a multiple of V_{i1}, reduction to the simple form as at (4.10) does not occur: more than the largest and the (new) smallest eigenvalues of $\{(\gamma Y_i' A_i)'\}$ are affected by seeking a rescaling matrix B for $\{(Y_i' A_i)'\}$.

One-factor Model (cont.)

Should the scale be determined by the component $\operatorname{cov}(A_i,\, V_{i1})$ in much the same way as V_{i1} is constructed via the vector $\mathbf{e}$? Or by the part of A_i belonging to the space containing the “signal”, i.e., the fraction $\operatorname{cov}(A_i, V_{i,n+1})/\sqrt{\operatorname{var}(A_i)\operatorname{var}(V_{i1})}$ of $\sqrt{\operatorname{var}(A_i)}$? Or by the whole of $\sqrt{\operatorname{var}(A_i)}$? Broadly speaking, the last possibility is closest to current ACT practice. The first possibility is consistent with Chapters 2 and 3, and effectively asserts that if $\{A_i\}$ is poorly correlated with $\{V_{i1}\}$ then the present scaling procedure spreads scores out too much relative to the spread of the averages. This implies a similar muddying operation to the one noted already below (4.5).

Put another way, the first possibility has the effect of asserting that, if school-based scores diverge markedly from whatever external ranking criterion is offered, then a conservative approach to scaling is adopted, with both above and below average scores compressed more towards the mean than at present. This is consistent with A 4 : if aptitude is measured validly by ASAT scores, then when $\{A_i\}$ and $\{V_{i1}\}$ show a larger than reasonable residual mean square on the scale of the ASAT scores, there is divergence from the "valid ASAT score measure", and this divergence should not be used to excess in promoting an unwarranted inflation in the spread of the school-based scores. This is also consistent with the analysis of discrepancies in Chapter 6, with being conducive to a reduced incidence of "outlier" scores of ASAT relative to the school-based scores, and with the consistency of using as a reference scale for $\{V_{i1}\}$ an optimal mixture of $\{(A_{iQ}\ A_{iV}\ A_{iW})\}$. Above all, incorporating scores $\{A_i\}$ into the one-factor model makes the estimation of $\{v_i\}$, whether from course scores or ASAT scores, a consistent exercise.

For all these reasons we choose this first approach.

Weight of ASAT Scores in a One-factor Model

There are two practical considerations involved in choosing a weight for ASAT scores in their contribution to estimates of the scaling criterion variable v_i. First, not all course scores X_{ij} are in fact scaled by whatever statistical scaling method is used, due to their coming from groups with small numbers of students. When such numbers are small, the error variables in the scaling criterion variables can easily dominate the scaling operation: all systems in Australia have evolved procedures to cope with "small groups", tantamount to giving greater weight to examiners' or teachers' experience in setting the level of the course scores concerned. In the ACT, procedures for fixing small group course scores also include reference to both the ASAT scores and other course scores of the students concerned.

REMARK 4.1. **The ACT system already uses Other Course Score scaling criteria.**

Fixing some scores independently of a statistical scaling process has two implications:

(1) these scores should have equality of input with other scores in determining the overall use of the reference scale scores if at all possible; and

(2) there is no reason *a priori* that such scores should affect systematically the signal to noise ratio of the scores at the college of the students concerned.

We deal with (1) by including the independently fixed scores in estimates of the student parameters v_i which are scaled against the reference scale. We can cope with (2) in either or both of two ways. One is to alter the scale factor (currently, 25.0) of the reference scale scores so as to yield TE scores of about the same spread as existing scores. The other is to choose a larger or smaller weight for ASAT scores so as again to yield TE scores of similar spread to the existing scores. The choice of a weight also arises directly in implementing the scaling procedure in that, in the notation of Chapters 2 and 3, the estimates of the scale parameter β_A should be close to 1.00. In practice, using the latter device on its own requires a weight of 1.5 course scores whereas applying the results of the algebra below to data indicates that the weight should be less than 1.0. Thus, a combination of the two methods is required. We consider now the question of what weight to attach to reference scale scores in a scaling procedure model, and defer the choice of an appropriate scale for ASAT scores in relation to small group scores to our discussion of ACT data in Chapters 10–12.

Use the one-factor model for scores

$$Y_{ij} = v_i + e_{ij}\,, \tag{4.21a}$$

$$A_i = v_i + e_{iA}\,, \tag{4.21b}$$

in which $\mathrm{var}(e_{ij}) = \sigma_j^2$, $\mathrm{var}(e_{iA}) = \sigma_A^2$. What weight w_A of reference scale scores in the estimator for v_i leads to (most nearly) unbiased estimators $\{b_j\}$ of the scale parameters $\{\beta_j\}$, and in particular b_A ? An explicit answer needs further constraints.

Typically, there is in place a prescription as to how scores Y_{ij} may contribute to an aggregate score, i.e., weights w_{ij} exist such that the aggregate is defined as

$$\sum_{j=1}^{n} w_{ij} Y_{ij}\,, \tag{4.22}$$

so that an estimator of v_i from scaled course scores alone is similarly defined as

$$\mathrm{est}(v_i) = \frac{\sum_{j=1}^{n} w_{ij} Y_{ij}}{\sum_{j=1}^{n} w_{ij}}\,. \tag{4.23}$$

By analogy therefore, inclusion of the reference scale scores is effected by using

$$\tilde{v}_i = \frac{\sum_{j=1}^{n} w_{ij} Y_{ij} + w_A A_i}{\sum_{j=1}^{n} w_{ij} + w_A} \tag{4.24}$$

as an estimator of v_i. Method of moment estimators of the scale coefficients are then

$$b_j = \mathrm{var}(\tilde{v}_i)/\,\mathrm{cov}(Y_{ij},\, \tilde{v}_i), \tag{4.25}$$

which includes the case $j = A$ on putting $Y_{ij} = A_i$. To be asymptotically consistent these estimators should be (asymptotically) 1.0, so we evaluate the expectation of numerator and denominator as

$$\mathrm{var}(\tilde{v}_i) = \mathrm{var}(v_i + \textstyle\sum_j w_{ij} e_{ij}/n_i) = \mathrm{var}(v_i) + \mathrm{var}(\textstyle\sum_j w_{ij} e_{ij}/n_i), \tag{4.26a}$$

$$\mathrm{cov}(Y_{ij},\, \tilde{v}_i) = \mathrm{cov}(v_i + e_{ij},\, v_i + \textstyle\sum_j w_{ij} e_{ij}/n_i) = \mathrm{var}(v_i) + \mathrm{var}(w_{ij} e_{ij}/n_i), \tag{4.26b}$$

with $n_i = \sum_j w_{ij}$, and the notation $\sum_j$ allows for different students to take different sets of courses. To make progress with (4.26), first fix the set of courses as a common set $j = 1, \ldots, n$ with $w_{ij} = 1$ for such j, so $n_i = n + w_A$. The equations become

$$\mathrm{var}(\tilde{v}_i) = \mathrm{var}(v_i) + \frac{\sum_j \sigma_j^2 + w_A^2 \sigma_A^2}{(n + w_A)^2}\,, \tag{4.27a}$$

$$\mathrm{cov}(Y_{ij},\, \tilde{v}_i) = \begin{cases} \mathrm{var}(v_i) + \sigma_j^2 (n + w_A) & \text{if } j \neq A, \\ \mathrm{var}(v_i) + w_A \sigma_A^2/(n + w_A) & \text{otherwise.} \end{cases} \tag{4.27b}$$

Substitute in (4.25). Algebraic simplification yields

$$b_j - 1 = \frac{w_A (w_A \sigma_A^2 - \sigma_j^2) + \sum_k (\sigma_k^2 - \sigma_j^2)}{(n + w_A)^2 [\mathrm{var}(v_i) + \sigma_j^2/(n + w_A)]} \tag{4.28a}$$

$$b_A - 1 = \frac{\sum_k (\sigma_k^2 - w_A \sigma_A^2)}{(n + w_A)^2 [\mathrm{var}(v_i) + w_A \sigma_A^2/(n + w_A)]}\,. \tag{4.28b}$$

From this last equation, the estimator b_A is model unbiased when

$$w_A = (\textstyle\sum_j \sigma_j^2)/n\sigma_A^2 \,. \tag{4.29}$$

With ACT data, it is commonly found that σ_A^2 is larger than (almost) all σ_j^2, so using w_A as at (4.29), the estimates of the scale parameters for the course scores are closer to unbiased as well. It is also evident from (4.28a) that if $\sigma_j^2 = \sigma^2$ for all courses j, then using (4.29) makes all the other estimators asymptotically unbiased like b_A.

Reference to the discussion around equations (4.1)–(4.4) shows that using (4.29) coincides with the implications of the bivariate adjustment approach underlying the argument there. This gives added support for our integrated one-step approach of treating the course scores $\mathcal{Y}$ and the reference scale scores as a single multivariate data set, rather than a two-step route with first estimates of v_i being extracted from the set of course scores and subsequently adjusted onto a "common scale" using a bivariate procedure with a separate set of reference scores.

Work of Cooney (1975, 1978) and Hasofer (1978) can also be read as studies in a bivariate setting of the question of finding a "natural" weight for reference scale scores. Within the framework of the concepts as above, neither Hasofer nor Cooney made plain that each component of the bivariate pairs of observations they considered constitutes an estimator of some common factor v_i with errors in both estimators. Cooney's use of a variance equating method as a resolution of the dilemma as to how to provide a scale is equivalent to assuming that both variables have the same measurement error variances, which in the present context means assuming that $\sigma_A^2 = \sigma^2 = \sigma_j^2$ for all courses j. However, the work is inappropriate for present applications because it concentrates on a bivariate rather than multivariate setting.

While any choice of weights w_{ij} is possible in (4.22), the choice as given is arguably the most appropriate in terms of reflecting the components of i's total studies: certainly it reflects existing practice in constructing both aggregates and scores themselves. Other possibilities include

(i) the use of weights that are inversely proportional to the measurement errors, and

(ii) the use of only those courses contributing to students' "best m subset" aggregates.

Possibility (i) assumes that measurement errors can be estimated satisfactorily, which presumes much more in the way of model robustness in what is being estimated than appears warranted on the evidence, especially in relation to the sizes of groups in the ACT where, except for English and Mathematics, almost all courses have around 100 or fewer student scores. The latter possibility would entail replacing v_i in (4.1a) by some larger quantity that depends on the number of courses that i takes (cf. the discussion on order statistics in Daley (1985)). It would also complicate applicability of the model, with selection effects dominating considerations and making invalid the use of the quantities v_i in estimating the parameters (α_j, β_j) in courses which do not contribute to i's aggregate. It may also encourage students to adopt a fragmentary approach to their curriculum and concentrate on a minimal number of courses required to contribute to an aggregate rather than studying their chosen curriculum more evenly (or as evenly as any of us may be wont!). In the context of the two-factor model it would imply the use of a quantity influenced more by the contrast factor v_{i2}.

Conclusion 4.4. **Reference scale scores and course scores should be treated as a single data set for scaling purposes using the Method-of-Moment estimation procedure to find the scale parameters $(\beta \; \beta_A)$ with a weight factor w_A attached to the reference scale scores given by (4.29), and the normalization constant for $(\beta \; \beta_A)$ determined by setting $\beta_A = 1$.**

Conclusion 4.5. **The multivariate data set approach of Conclusion 4.4 coincides with the two-stage approach of constructing school-based estimates of relative general**

achievement within a college, and subsequently combining these optimally with an external set of reference scores to produce system-wide estimates of relative general achievement.

Conclusion 4.6. **The one-step procedure of the Method of Moment estimation procedure can be dissected so as to furnish colleges, if they so desire, with approximate "within-college" estimates of their students' relative general achievements or "within-college" TE scores, at any time that a school-based set of quasi-course scores is known. In particular, a purely school-based set of relative achievement scores can be furnished, but such scores would have no between-school comparability. Such estimates would be subject to adjustment to reflect system-wide comparability at a final stage when system-wide reference scores are determined and made known.**

Educational Measurement Approach

The approach of educational measurement advisors in the ACT and Queensland as to how to use statistical methodology in a scaling procedure has been to construct a common assessment task (or set of tasks), administer it to all students, and construct a reference scale or scales from the resulting score(s) to be applied in a bivariate setting via a two-moment equating method. By this, I mean for example that the reference scores $\{R_i\}$ would be used to determine a scale for $\{V_{i1}\}$ for a given school (hence, a given set of student indices i) by means of the relations

$$\operatorname{ave}(R_i) = \operatorname{ave}(V_{i1}), \qquad \operatorname{var}(R_i) = \operatorname{var}(V_{i1}). \tag{4.30}$$

At present in the ACT, three scores are available for use as reference scores, namely the Quantitative and Verbal sub-scale scores of ASAT, and a Writing Task score. It has largely been an act of faith that these three scores, $\{Q_i\}$, $\{V_i\}$ and $\{W_i\}$ say, when combined in a certain prescribed fashion, constitute an adequate reference scale. While correlations are regularly calculated and published annually in the *Year 12 Study* and briefer information in Queensland, there has been relatively little discussion as to whether or not the scaling procedure is optimal, let alone that it is accomplishing the task it has been assumed to be executing fairly. And as reading the discussion in Masters and Beswick makes plain, when there is no model to provide any point of reference, the discussion can rapidly lose direction and relevance.

The Outlier Problem

Suppose we represent purely school-based measures $\{v_i\}$ and ASAT scores $\{A_i\}$ on comparable scales. Inspection of ACT data shows that for some few students, typically around 1% to 3 or 4%, the absolute differences $|A_i - v_i|$ exceed 3 times the standard deviation of the differences. These rates of difference exceed the Gaussian distribution rate of *c.* 0.1% by a large amount, being closer to the order of the upper bound of 5% of the Camp-Meidell inequality for unimodal distributions. Conversing with teachers in their franker moments leads to describing occasional students being "lazy loafers", i.e., higher academic *ability* and low *achievement* as determined by school-based assessment tasks. The unquestioned use of the ASAT score data of such individuals in any scaling procedure that assumes their equivalence as at A 5 with course scores to within the limits of measurement error, and hence assumes the similarity of achievement and ability measures, introduces biases and unnecessary noise into an already potentially noisy operation.

It is also the case that ASAT scores of students for whom A_i is in the "nonsense" score region corresponding to an average score obtained by random guessing (below about 30% "correct" items) are used without question.

While we have shown at (4.29) how to fix a weight for using ASAT scores of most students, the practical problem of coping with students whose scores do not meet the criteria that make any statistical scaling operation valid, has been shunned by both the Accrediting Agency and the 1986 Review Committee. "She'll be right" attitudes do not apply to the use of statistical procedures which can be quite sensitive to departures from assumptions, as ACT experience has amply demonstrated. It is the case that allowance is made for certain groups of students, identified by external criteria, whose ability and achievement measures may not necessarily match up as groups: for NESB and MA students the use of their ASAT scores is annulled by *prescriptive* fiat. What is needed is the extension of this idea to include some statistical quality control to ensure that the ASAT and course scores used do more consistently measure some common underlying factor. The existing practice of using scores of glaring outliers is a wanton disregard for the practicalities of applying a statistical technique in such a precise fashion as occurs.

Thus, just as Masters and Beswick commented about lack of a statistical model, so the following should not be surprising.

CONCLUSION 4.7. **The present statistical use of ASAT scores lacks attention to critical detail. Scrutiny shows that the data deviate excessively from the assumed close relationships. Without these relationships, further checks and adjustments are necessary to justify the statistical use of ASAT scores.**

Several Reference Scales?

In discussion in 1986 following completion of *MATHEF* it was decided to use more than one reference scale in the scaling of course scores to produce ACT TE scores. This step involves some illogical thinking, and biases ("statistical artefacts") can result.

For example, the motivation for using ASAT-Q scores as a reference scale for Mathematics course scores comes primarily from face validity considerations. These are supported by statistical analyses which typically show that

$$\text{corr}(\text{Maths}, Q) > \text{corr}(\text{Maths}, T) > \text{corr}(\text{Maths}, V). \tag{4.31}$$

However, it is typically also the case that

$$\text{corr}(\text{Maths, mean of Other Course Scores}) > \text{corr}(\text{Maths}, Q), \tag{4.32}$$

so on statistical grounds, means of Other Course Scores provide an even better reference scale. In modelling terms, it is certainly the case that the set of scores having the maximum correlation is "best" because the space complementary to the direction of the first principal component ("common factor") is then smaller and of lesser possible influence, and it is in this complementary space that uncertainties and biases are found. In the case of (4.31)–(4.32), the same set of Mathematics course scores is involved, so in representing them as

$$M_i = \mu_i + e_{iM}\,, \tag{4.33a}$$

$$(\text{Ref'ce scale})_i = \mu_i + e_{i,\text{Ref}}\,, \tag{4.33b}$$

the inequalities concerned are equivalent to the error variances $s^2_{\text{Ref}} \equiv \text{var}(e_{i,\text{Ref}})$ satisfying

$$s^2_{\text{OCS}} < s^2_Q < s^2_T < s^2_V\,. \tag{4.34}$$

If we judge scaling criteria by "small errors", then means of Other Course Scores yield the best.

Inspection of data reveals that students whose scores most influence these errors (those at the extremes of the Mathematics scores) tend to be those with more (higher end) or fewer (lower end) science course scores. This implies that there is also some face validity in the use of Other Course Score scaling criteria: it is not just a piece of arithmetic.

In systems like NSW which use both examination and school-based marks, the examination marks currently serve two purposes: they are assessment marks for certification purposes, and they provide a system-wide reference scale for scaling each school's set of school-based assessments in the course concerned. This enables the latter to be reported on the Certificate also in terms of being comparable with the examination-based score. These exam.-based scores have a mean of 60 and standard deviation of 12.5 for the candidature of each 2-Unit subject. For the purposes of producing a scaled aggregate for admission to a university through UCAC or to the Canberra CAE, the average of the two scores is used after being scaled by an Other Course Score scaling procedure to adjust for the different candidatures in the various courses.

The use in the ACT of sub-scale scores in 1986–88, and in Morgan & McGaw's (1988) study of the possible use of a variety of reference scale scores, in part reflects a desire to emulate a family of separate reference scales as though derived from external examinations, without the curriculum constraints entailed in students being required to take such exams. It has been *presumed* that standardizing these reference scales to the same mean and standard deviation over a common population, is enough to ensure "fairness" in their use. We shall see empirically in Chapter 12 (e.g. Tables 12.1, 12.2 and 12.7) that this need not be the case.

In modelling terms this departure from "fairness" can be seen on two grounds as follows. Take two pairs of course scores, each with a different reference scale, the latter scores having the same first two moments for a common population. Suppose that for one pair of course and reference scale scores the error variance is rather smaller than for the other pair (think for example of Mathematics with Q, and English with V or T). Then firstly, what is being substantively measured in terms of (4.33) has a lesser spread where the error variance is higher (recall McGaw, 1987). Equivalently, in terms of the analysis of Chapter 2, course scores do not generally contribute "equally" to an aggregate merely by virtue of having the same mean and standard deviation. Secondly, it is likely that students who score more highly on the more specific reference scale will be concentrated in that course rather than the other. This means that by choice of course, some students can be guaranteed a higher mean reference scale score when there are several fixed reference scales being used rather than a single scale (it is this effect that shows visibly in Table 12.7).

Another aspect of the selection effect of skills required for various courses, can be based on a suggestion that skills are developed in students in relative terms, and that measures like ASAT scores basically rely on these exceeding some base level. For students not satisfying this assumption, a base level or "nonsense" ASAT score results. Current educational pressures tend to encourage students to choose science related courses if their skills are sufficiently above the assumed base level. Consequently, students opting for those courses whose scores are scaled against ASAT-Q in the ACT, tend to be selected as those with higher ASAT scores, and whose ASAT-Q scores are higher still.

Conclusion 4.8. **In terms of both face validity and modelling considerations, Other Course Score scaling is both consistent with a model aimed at producing a single aggregate and with yielding an aggregate free(r) of bias from selection effects.**

Chapter 5

Some Statistical Properties of ASAT Scores

The object of this chapter is to provide some information on statistical properties of ASAT scores. The particular properties that concern us relate to their use as reference scales in Queensland and the ACT. This essentially means asking how consistently such scores or sub-scale scores can be represented (cf. equation (4.1b)) as

$$A_i = v_i + e_{iA}\,. \tag{5.1}$$

Measurement Error Variances

Every year the Australian Council of Education Research supplies certain information to the users of ASAT on some basic properties of the ASAT scores for the version used in testing towards the end of the previous year. I start by discussing the measurement error variance τ_A^2.

Up to and including the 1985 test τ_A^2 was determined directly from the K–R 20 formula[11] . This formula is deduced by assuming that each student's responses to the questions or *items* on a test of length N have statistically independent errors (i.e., choosing incorrect responses to the various items). When applicable, the formula implies that τ_A^2 cannot be any larger than $N/4$. These independence assumptions are quite strong, and at face value certainly of dubious applicability to ASAT papers where the items are grouped together, referring to a common passage of stimulus material (i.e., reading matter to be comprehended and providing a context for the questions). In 1986 an alternative estimate of τ_A^2 was computed by the split-half method: since in the ACT and Queensland each year's ASAT paper is administered as two papers each of 50 items, and these papers are so compiled as to be of equivalent difficulty and scope of material, a measure of the reproducibility of the scores is furnished by comparing total scores on the two separate papers. In the ACT, Queensland and WA this yielded consistent estimates of $\tau_A^2 \approx 30 \approx 5.5^2$ rather than the estimates *c.* $20 \approx 4.5^2$ obtained from the K–R 20 formula over several years. In what follows, I use this larger figure as a measure of τ_A^2 for an ASAT total score, denoted here as ASAT-T to distinguish it from the sub-scale scores ASAT-Q and ASAT-V.

In Queensland, the raw total score is rescaled for use as ASAT-T for scaling purposes. In 1985, the standard deviation of the crude scores[12] was 14.2, so if we can relate τ_A^2 to e_{iA} in (5.1) by

$$\tau_A^2 \equiv \operatorname{var}(e_{iA})\,, \tag{5.2a}$$

[11] So named from its appearance as equation (20) in a paper by Kuder & Richardson (1937)!

[12] The data concerning 1985 ASAT scores and its sub-tests are taken from Appendix 3 of *Tertiary Entrance in Queensland: A Review* (1987). As an example of data from another year—and I thank ACER for this information—the correlation coefficient of scores on the two papers equals 0.763, with standard deviation of the Total raw score 14.018, and of sub-scale scores 9.154 and 6.450 for Q and V respectively, these being based on 55 and 41 items. The measurement error variance of T is then

$$14.018^2/[1 + 2(0.763/0.237)] = 26.416 = 5.140^2.$$

With the same assumptions as for (5.3), those entries are replaced by

$$(55/100) \times 26.42 \times (25/9.154)^2 = 108.4 \quad \text{and} \quad (41/100) \times 26.42 \times (25/6.450)^2 = 162.71.$$

Thus the measurement error variance in 1985-style ACT ASAT-T is $1.10^2 \times (108.4 + 162.7)/4 = 82.00$, which is smaller than the whole test figure of $(25/14.018)^2 \times 26.416 = 84.02$. The signal to noise ratio equals $(625 - 82.00)/82.00 = 6.62$, and error variance of $Q - V$ equals $271.1 = 16.46^2$.

then the "signal to noise" ratio

$$\mathrm{var}(v_i)/\tau_A^2 = (14.2^2 - 30)/30 = 5.72. \tag{5.2b}$$

In the ACT, ASAT-T was initially constructed as in Queensland. Starting *c.* 1984 it was constructed by rescaling ASAT-$Q$and -V sub-scale scores to a common standard deviation, forming the 50 : 50 average of these two scores, and rescaling the result so that the population standard deviation equals 25. (Onwards from 1986 a Writing Task score has been included in the Verbal component, but this does not concern us for the moment.) In 1985, the raw ASAT sub-scale scores Q and V were based on 49 and 47 items and had standard deviations of 8.9 and 6.4 respectively. Regard[13] the measurement error τ_A^2 as being additive over and proportional to the number of items, so the contribution per item equals 0.3, and the measurement error variances of raw Q and V are 14.7 and 14.1 respectively. Rescaling the sub-scale scores to a standard deviation of 25.0, these increase to

$$14.7 \times (25/8.9)^2 = 116.0 \quad \text{and} \quad 14.1 \times (25/6.4)^2 = 215.1. \tag{5.3}$$

Take corr$(Q,\ V) = 0.65$. Then to rescale the average $(Q+V)/2$ to a standard deviation of 25.0 we must multiply it by $\sqrt{2/(1+0.65)} = 1.10$. Using independence of the measurement errors again, this implies that the measurement error in the 1985 ACT ASAT-T score, when it has a standard deviation of 25, equals

$$1.10^2 \times (116.0 + 215.1)/4 = 100.3, \tag{5.4}$$

(cf. $(25/14.2)^2 \times 30 = 93.0$ had the raw ASAT-T score been used). This slightly larger figure implies that for the ACT in 1985, when τ_A^2 in (5.1) is taken as the measurement error of A_i,

$$\mathrm{var}(v_i)/\tau_A^2 = (25.0^2 - 100.3)/100.3 = 5.23. \tag{5.5}$$

Later we shall need an estimate of the measurement error of $Q - V$: for 1985 this equals $4/1.10^2$ times the error in ASAT-T, i.e., $3.30 \times 100.3 = 331.6 = 18.2^2$.

Recall that the standard deviation of the raw scores on a test of length N is to the first order proportional to N. It is not uncommon for the number of items used in constructing the verbal sub-scale scores to be rather less than the number in Q. If we now adjust the existing figures for a test in which the numbers of items used are 52 for Q and 42 for V (and 6 not included because the responses are either inconsistent or else fail to distinguish between the Q and V sub-scales), we should have in place of (5.3)

$$15.6 \times (25/[8.9 \times 52/49])^2 = 109.3 \quad \text{and} \quad 12.6 \times (25/[6.4 \times 42/47])^2 = 240.8,$$

giving as a measurement error in a 1985-style ASAT-T, in place of (5.4),

$$1.10^2 \times (109.3 + 240.8)/4 = 105.9.$$

The situation in the ACT is now a little more complex as Writing Task scores W are also used. The raw scores for the Writing Task are compiled by adding four readers' scores, each on the scale $\{0, 1, \ldots, 6\}$ (I understand that the score 0 has never been given, so for norm-referencing purposes the scale is quite coarsely graduated); raw scores are therefore in the range 4 to 24. The actual distribution in 1986 is shown in Table 5.1; it has mean 12.72 and standard deviation 3.30.

[13] The assumptions of additivity and constancy of the components of the measurement error variance are certainly too strong, else the Kuder–Richardson formula should be applicable! Simply as an approximation, the likely effect for usage below is that the measurement errors of the sub-scale scores are over- and under-estimated for Q and Vrespectively.

TABLE 5.1

Distribution of Raw Writing Task Scores in 1986

Score	# F	# M	All	Score	# F	# M	All
4	2	11	13	15	122	83	205
5	9	20	29	16	121	65	186
6	9	21	30	17	67	46	113
7	17	45	62	18	56	28	84
8	30	61	91	19	39	11	50
9	60	86	146	20	19	5	24
10	93	119	212	21	7	1	8
11	97	121	218	22	1	2	3
12	150	118	268	23	2	0	2
13	159	137	296	24	1	0	1
14	154	81	235				

Because the standard deviation of the raw scores is smaller than 4, an appreciable component of any measurement error is attributable to the discrete nature of the scale. Scores equal to (say) 14 necessarily come from a set $\{3, 3, 4, 4\}$ (or maybe even $\{3, 3, 3, 5\}$ etc.) so if a true score model is appropriate, their measurement error variance is at least 4.0. There is little information from which to estimate the overall measurement error variance. There are two obvious components of this variance: one is from the discrete nature of the scale (to me, it seems too coarse), while the other is due to the essay topic — and since only one essay is written each year, direct estimation of error from that source can only be by inference from elsewhere. If we guess that the measurement error variance equals 3.0, then the signal to noise ratio of the test is about 2.63.

The computation just given is predicated on an application of the true score model to individual scores. It is arguable that by assessing the Writing Task subjectively four times, breadth of assessment criteria from different examiners is provided so that e.g. scores of 3 and 4 from different examiners are giving different information with no scores being "incorrect". Since the total mark is constructed by aggregation of marks from the same four examiners, their sum defines a ranking via the same mix of assessment criteria.

As for information from "elsewhere", a crude idea of measurement error on a single essay (but, written at speed rather than in a more considered fashion in response to much stimulus material) can be gleaned from published NSW HSC data in the 1-Unit General Studies course for which a student is required to write four essays in the examination. For the three years 1984–86 the correlation r_{ES} say between exam. marks and school-based assessments was observed to lie in the range 0.77 to 0.79. Now a true score model shows readily that the signal to noise ratio for the four exam. essays equals $1/[(1/r_{\mathrm{ES}}) - 1]$, assuming that school-based assessments and exam. marks have equal "errors", or else $1/[(1/(r_{\mathrm{ES}})^2 - 1]$ if the school-based assessments are regarded as having much smaller errors than the exam. paper: the former suits present purposes better in the sense of leading to a larger ratio (!). This signal to noise ratio SNR say, is for four essays: the ratio for a single essay, assuming independence of "errors" in the scores on each essay, equals $SNR/4$. For $0.77 < r_{\mathrm{ES}} < 0.79$, this yields $1.09 < SNR/4 < 1.19$: this is rather smaller than our guesstimate for the Writing Task scores.

CONCLUSION 5.1. **The units of the integer-valued sub-scale scores Q, V and W corresponded in 1986 to 12%, 17% and 30% of their respective standard deviations. The measurement error standard deviations of these scores are about 4 units for Q and V and 2 to 3 units for W.**

TABLE 5.2

Gender-linked Differences of ASAT Sub-scale Scores 1979–85 and their Standardized Gender-linked Discrepancies.

Year	$Q_F - Q_M$	$V_F - V_M$	$- D_{FM}(Q, V)$		
	ACT	ACT	ACT	WA	Qld
1979	– 4.0	0.4	12.6	14.6	14.3
1980	– 3.1	0.0	8.6	12.7	10.8
1981	– 5.6	– 0.4	14.0	15.4	14.8
1982	– 5.2	– 1.1	10.2	14.3	12.0
1983	– 4.1	0.2	12.2	9.2	12.5
1984			14.0		
1985			9.8		

Note: Differences of sub-scale scores Q and V are of raw scores, so they are rescaled to standard deviation of at most 25.0 by multiplication by 25/9.0 and 25/6.5 respectively, so for 1979 for example, $-12.6 = (25/9.0)(-4.0) - (25/6.5)(0.4)$. 1979–83 data are from Table 3.1 of Adams (1984), 1984–85 data from Masters & Beswick (1986).

Psychometric Stability of ASAT Scores

The second question we consider is the psychometric stability of ASAT scores. On the basis of published evidence this question can be addressed in at least two ways. First, ask whether the gender-linked difference $D_{FM} \equiv A_F - A_M$ of mean ASAT scores for females and males is stable between each paper (a new paper is produced each year). Three independent pieces of information were cited in Daley (1985) indicating that in this respect the papers are noticeably unstable:

(i) regression analyses of the differences D_{FM} that remove the effect of different retention rates show too large a residual variation;

(ii) the differences D_{FM} in the three regions using the test (WA, Queensland, and the ACT) are more consistent between different regions for a given paper than they are between papers for a given region;

(iii) the analyses of gender-linked discrepancies between ASAT and school-based assessments (TE scores) in the ACT are distinctly different between different papers.

Second, sub-scale scores are produced for each test, and we can then find the gender-linked discrepancy between such scores. Write Q_F, Q_M, V_F, V_M for the sub-scale means, so that (approximately), when the overall mean score is standardized to 0 and the numbers of females and males are approximately equal,

$$Q_F \approx -Q_M\,, \qquad V_F \approx -V_M\,,$$

whence for the gender-linked discrepancy we have

$$D_{FM}(Q,\,V) \equiv Q_F - Q_M - V_F + V_M \approx 2(Q_F - V_F)$$

which has measurement error variance $4 \times 18.2^2/N_F$ (see the note below (5.5)). In the ACT $N_F \approx 1000$ so the standard error of $D_{FM}(Q,\,V)$ is about $\sqrt{36.4^2/1000} \approx 1.32$.

The ACT gender-linked discrepancies in Table 5.2 yield a chi-square test statistic of $27.0/1.32^2 = 15.5$ on 6 d.f., which exceeds the 95% significance level. The estimates for 1979–83 are if anything under-scaled, so the test statistic is on the conservative side.

Our next conclusion comes firstly from (i)–(iii). It is reinforced in the analysis as just described, from which we also conclude that the Quantitative and Verbal sub-scale scores are not psychometrically stable across the ACT population. Note that if we computed the corresponding statistic for Queensland, the larger population (upwards of 20,000) would yield a much smaller standard deviation in place of 1.32 , e.g. the chi-square test statistic is about $10.9/(36.4^2/10000) = 82.3$.

CONCLUSION 5.2. **For the purpose for which ASAT scores are used in the ACT, neither they nor the Quantitative and Verbal sub-scale scores are psychometrically stable with respect to gender differences in different years.**

Relative Structure of Quantitative, Verbal and Writing Task Scores

In Tables 5.3 and 5.4 the correlations of the three reference scale scores are shown for the whole system, also broken down by sex, and for most TE-score qualified students at each college. In detail, the scores used in Table 5.4 exclude those of non-English speaking background and Mature Age students; Table 5.3 was computed with Tables 6.1–2, where by requiring both English and Mathematics scores as well, about another 10% of students were excluded. The differences between these subsets and the whole population do not have a significant impact on the thrust of our conclusions.

In comparison with other studies (e.g. Breland & Griswold, 1982) the surprising feature here is that corr(V, W) is not higher, closer to 0.5 say, especially given that corr(Q, V) is around 0.65. Interpret 0.65 as indicating that a broad range of general ability is represented. In terms of face validity, one would expect the ASAT-V sub-scale scores to be closer to the Writing Task scores than the $-Q$ sub-scale scores. After all, these two sub-scales are constructed by looking for both divergence between responses to the different items on the ASAT paper and convergence towards the humanities and science sets of questions respectively, *i.e.*, they are deliberately made to be “different”. The interpretation[14] of the correlations is that, contrary to expectations based on face validity, the ASAT-V sub-scale scores are closer to the $-Q$ sub-scale than the Writing Task scale, an interpretation also borne out by the work in Chapters 4, 6 and 7. In addition to different measurement error properties, the reason may also be tied in with the contrast in modes of assessment between the multiple choice scores Q and V and the essay format of the Writing Task: the former emphasizes problem solving skills while the latter is much freer, with selection of material left to the student.

TABLE 5.3

Correlations of Q, V and W scores within each college

College	Q/V	Q/W	V/W
1	0.577	0.215	0.336
2	0.667	0.279	0.346
3	0.483	0.243	0.367
4	0.550	0.355	0.416
5	0.656	0.368	0.432
6	0.689	0.329	0.420
7	0.634	0.111	0.385
8	0.634	0.292	0.356
9	0.679	0.426	0.356
a	0.652	0.489	0.368
b	0.659	0.210	0.338
c	0.712	0.460	0.483
d	0.677	0.313	0.290

[14] But, note as before that our conclusion is based on just one ASAT paper and Writing Task.

TABLE 5.4

Correlations and Covariances of Writing Task and ASAT Sub-scale Scores and their Loadings

	Correlations			Covariances		
	(a) All Students (excl. NESB and Mature Age)					
Q	1.000			1.010		
V	0.626	1.000		0.633	1.013	
W	0.249	0.387	1.000	0.250	0.390	1.000
Q Loading	0.824	0.418	0.382	0.830	0.416	0.384
V Loading	0.883	0.159	– 0.441	0.891	0.156	– 0.043
W Loading	0.634	– 0.764	0.117	0.631	– 0.767	0.119
Latent Roots	1.863	0.783	0.354	1.879	0.786	0.358
% Var. Explained	62.1%	26.1%	11.8%	62.2%	26.0%	11.8%
	(b) Female Students Only					
Q	1.000			1.025		
V	0.651	1.000		0.644	0.956	
W	0.316	0.385	1.000	0.305	0.359	0.907
Q Loading	0.847	0.354	0.396	0.876	0.338	0.378
V Loading	0.876	0.221	–0.428	0.858	0.169	–0.437
W Loading	0.659	–0.750	0.061	0.600	–0.736	0.073
Latent Roots	1.919	0.737	0.344	1.865	0.684	0.339
% Var. Explained	64.0%	24.6%	11.5%	64.6%	23.7%	11.7%
	(c) Male Students Only					
Q	1.000			0.835		
V	0.664	1.000		0.631	1.081	
W	0.353	0.404	1.000	0.322	0.419	0.993
Q Loading	0.854	0.339	0.394	0.754	0.291	0.427
V Loading	0.876	0.238	–0.420	0.927	0.298	–0.364
W Loading	0.682	–0.730	0.046	0.691	–0.717	0.023
Latent Roots	1.962	0.704	0.333	1.906	0.688	0.315
% Var. Explained	65.4%	23.5%	11.1%	65.5%	23.7%	10.8%

Chapter 6

Discrepancies Between ASAT and Course Scores

This chapter overlaps in part with Chapter 7, but is closer in spirit to Chapter 5, in pursuing properties of ASAT scores *per se* in relation to course scores. The more specific matter of the "sex bias in ASAT" problem is noted, though not comprehensively, in Chapter 7.

Gender-linked Mode of Assessment Discrepancy

Ever since the end of 1978, if not twelve months earlier, it has been known[15] in the Australian Capital Territory that there exists a gender-linked discrepancy between ASAT and course scores. The type of discrepancy observed[16] has been known in educational psychology and educational measurement circles for decades, though it has not necessarily been recognized as being associated with the use of multiple choice methodology for assessment as opposed to those methods of assessment like essay-writing commonly used in external examinations and school-based assessments, especially but not exclusively outside of the so-called exact sciences. Now that suitable data are available, the hypothesis has been tested in more detail as below as a result of which it follows that the discrepancy is properly referred to as being *gender-linked* on account of the last part of the following statement.

[15] Immediately after the first ACT Tertiary Entrance scores were issued in 1977, some girls' schools contacted the ACT Schools Accrediting Agency. They queried these new scores on the grounds that the proportion of their students eligible for admission to the Australian National University on the basis of this new aggregate score was significantly decreased from the proportion eligible in 1976 and earlier on the basis of New South Wales HSC aggregate score (see Table 7.1). At the same time, there was a marked increase in the proportion of eligible students from boys' schools (this change may have been noted, but to my knowledge it was not queried (!)). Twelve months later, Morgan (1979) analysed results from December 1978 and observed a gender-linked discrepancy between ASAT and TE scores. These analyses were repeated independently in 1984 (see Daley, 1985), in ignorance of Morgan's work. In mid-1986, after the report *MATHEF* had been written, a fuller data set became available to document the 1977 query (see Table 7.1). All these data made it plain that the reason for the shift in 1977 was a sex bias in ASAT. There has been a lamentable failure, certainly since 1984, if not before, to remove the effects of the bias. On the basis of all the evidence available, statistical adjustment of the somewhat variable ASAT scores is the surest way of producing TE scores that best reflect the policy principles P 1–4 concerning school-based assessments. For reasons why the bias still persists, see §1.19 of *MATHEF* for some formal reservations, and footnotes 2 and 3 to Chapter 7.

[16] Daley (1986a) has a literature review covering experiences in USA and UK, including external examination systems in UK. To this can now be added (i) the detailed analyses summarized in Tables 6.1–3 below; (ii) Writing Task *v.* ASAT multiple choice Verbal scores for 1986 and 1987 as in Table 6.5; (iii) data documenting the original complaint in Table 7.1; and (iv) data from Queensland similar to those of Daley (1984) (these data consist of 1987 ASAT scores and the Queensland internally produced RAG score, that system's analogue of the ACT TE score: see Figure 3 of *ASAT and TE Scores* (1988)). No official report contains anything close to a comprehensive account of the variety of evidence available: *MATHEF* gave a biased account of the matter by failing to note significant information that ran contrary to its preferred conclusion: it mentions neither the ASAT sex difference variability nor the consistent gender-linked discrepancies in quantitative and verbal domains (see Table 6.3).

CONCLUSION 6.1. **Students' relative abilities as assessed by multiple choice methods in the ASAT test and reported on Quantitative and Verbal sub-scales, are positively correlated with but differ systematically from school-based determinations of their relative achievements in the related areas of Mathematics and English respectively, this difference being a characteristic of the two modes of assessment used in the two pairs of scores. Irrespective of the Quantitative or Verbal "dimension" concerned, females tend to perform relatively better on the school-based measures and males on the ASAT test.**

The evidence for Conclusion 6.1 is given in Tables 6.1 and 6.2. It is based on the scores as used and/or reported on 1986 Year 12 Certificates, i.e., the school-based data have been scaled by the existing procedures. Using scaling parameters estimated by the Method of Moments would make little difference to the thrust of the analyses.

We start by illustrating the use of the terms gender-linked *difference*, gender-linked *discrepancy*, and gender-linked *bias*, in the context of average scores A_F and A_M for females and males on the ASAT paper, and similarly Y_F and Y_M in some course. Each pair of scores yields a *gender-linked difference*, namely $A_F - A_M$ and $Y_F - Y_M$. We call their difference $(A_F - A_M) - (Y_F - Y_M)$ the *gender-linked discrepancy* of the two scores or measures; it must be approximately zero for the scores to measure a common factor subject only to measurement error, as is required of reference scores for a "fair" scaling operation (cf. also Chapter 8). If the discrepancy is systematically positive or negative, there is then evidence of a significant non-zero gender-linked discrepancy between the two purportedly similar measures. When e.g. a policy declaration identifies one measure as being the "correct" measure to use, we speak of a *gender-linked bias* in the other measure when there is a significant non-zero gender-linked discrepancy between the two measures.

CONCLUSION 6.2. **On the basis of the TE score construction principles P 2–4, Conclusion 6.1 implies that ASAT scores are biased for use as reference scores in the ACT. If similar principles hold in Queensland, the same conclusion holds there also.**

Denote student i's scores in English, Mathematics, and ASAT-Q and -V sub-scales by E_i, M_i, Q_i, V_i respectively. Regard these scores as being represented by expressions of the form

$$\begin{aligned}\text{(score)} &= \text{(general ability } \textit{cum} \text{ achievement)} + \text{(quantitative/verbal contrast)}\\ &\quad + \text{(school-based assessment/ASAT multiple choice contrast)} + \text{(error)},\end{aligned} \tag{6.1}$$

or in algebraic notation,

$$E_i = v_i + v_{i2} + \Delta_i + e_{iE}\,, \tag{6.2a}$$

$$M_i = v_i - v_{i2} + \Delta_i + e_{iM}\,, \tag{6.2b}$$

$$V_i = v_i + v_{i2} - \Delta_i + e_{iV}\,, \tag{6.2c}$$

$$Q_i = v_i - v_{i2} - \Delta_i + e_{iQ}\,, \tag{6.2d}$$

where the student parameters v_i, v_{i2}, Δ_i, have variances S_1^2, S_2^2, S_3^2 respectively, and the four error terms we shall for the moment regard as being mutually uncorrelated and uncorrelated with the student parameters, and having the same variance s^2. Simple algebra then leads to expressions for the three student parameters and an overall error term as

$$v_i + e_{i1} = \tfrac{1}{4}(E_i + M_i + V_i + Q_i), \tag{6.3a}$$

$$v_{i2} + e_{i2} = \tfrac{1}{4}(E_i - M_i + V_i - Q_i), \tag{6.3b}$$

$$\Delta_i + e_{i3} = \tfrac{1}{4}(E_i + M_i - V_i - Q_i), \tag{6.3c}$$

$$e_{i4} = \tfrac{1}{4}(E_i - M_i - V_i + Q_i), \tag{6.3d}$$

where the error terms represent

TABLE 6.1

Summaries of Simple ANOVA on English, Mathematics, ASAT-Q and -V scores

No.	College	Means			Standard Deviations			
		Sum	QVDiff	MADiff	Sum	QVDiff	MADiff	Error
			(a) All Students					
175	1	–4.565	–2.009	–0.678	19.006	9.613	9.450	6.773
169	2	1.274	–0.201	–0.109	20.329	8.689	8.376	6.332
212	3	2.892	–1.564	–0.919	16.796	9.142	9.006	6.089
79	4	–0.424	–2.393	–1.253	17.747	9.393	9.773	6.705
244	5	1.899	–0.020	–0.515	20.489	9.959	10.202	6.719
178	6	5.144	–1.070	–1.278	19.538	9.021	9.140	6.378
174	7	6.304	–1.224	–0.429	17.824	9.720	9.338	5.831
112	8	1.188	–1.587	–1.018	18.291	8.797	9.137	6.046
110	9	0.687	1.958	0.324	19.179	8.099	8.507	5.461
56	a	–5.576	4.808	0.785	19.388	8.151	8.646	5.897
116	b	–0.056	6.321	1.170	19.433	8.435	6.648	6.338
72	c	0.578	–3.574	–0.077	20.133	8.173	9.463	5.723
120	d	1.047	–4.531	–0.339	19.228	7.965	8.264	5.610
			(b) Female Students Only					
95	1	–7.618	1.588	1.828	17.858	8.489	8.807	6.998
78	2	–2.180	3.949	1.487	18.784	7.218	8.675	6.108
114	3	4.690	0.706	1.089	17.294	7.616	9.029	6.246
37	4	–0.617	1.506	2.438	18.353	9.803	7.580	5.253
124	5	0.681	3.528	2.339	20.401	8.562	9.713	6.283
91	6	7.181	2.179	0.727	19.674	7.130	8.508	6.048
88	7	5.113	2.695	1.716	16.509	9.018	9.243	6.040
61	8	–2.281	1.270	2.059	15.989	8.253	7.953	5.778
110	9	0.687	1.958	0.324	19.179	8.099	8.507	5.461
56	a	–5.576	4.808	0.785	19.388	8.151	8.646	5.897
116	b	–0.056	6.321	1.170	19.433	8.435	6.648	6.338
			(c) Male Students Only					
80	1	–0.939	–6.281	–3.655	19.790	9.152	9.375	6.540
91	2	4.235	–3.758	–1.476	21.221	8.285	7.905	6.549
98	3	0.799	–4.205	–3.254	16.031	10.056	8.438	5.933
42	4	–0.255	–5.827	–4.504	17.417	7.594	10.398	7.792
120	5	3.158	–3.685	–3.465	20.588	10.004	9.887	7.146
87	6	3.013	–4.468	–3.376	19.276	9.561	9.353	6.739
86	7	7.523	–5.234	–2.623	19.097	8.769	8.965	5.593
51	8	5.337	–5.005	–4.698	20.091	8.259	9.169	6.307
72	c	0.578	–3.574	–0.077	20.133	8.173	9.463	5.723
120	d	1.047	–4.531	–0.339	19.228	7.965	8.264	5.610

$$e_{i1} = \tfrac{1}{4}(e_{iE} + e_{iM} + e_{iV} + e_{iQ}), \tag{6.4a}$$

$$e_{i2} = \tfrac{1}{4}(e_{iE} - e_{iM} + e_{iV} - e_{iQ}), \tag{6.4b}$$

$$e_{i3} = \tfrac{1}{4}(e_{iE} + e_{iM} - e_{iV} - e_{iQ}), \tag{6.4c}$$

$$e_{i4} = \tfrac{1}{4}(e_{iE} - e_{iM} - e_{iV} + e_{iQ}), \tag{6.4d}$$

and by assumption they are mutually uncorrelated with common variance $s^2/4$.

TABLE 6.2

Components of Variance from Table 6.1

College	Standard Deviations			
	Sum	QVDiff	MADiff	Error
	(a) All Students			
1	17.758	6.822	6.590	6.773
2	19.318	5.950	5.483	6.332
3	15.653	6.819	6.636	6.089
4	16.432	6.578	7.110	6.705
5	19.356	7.351	7.677	6.719
6	18.468	6.380	6.547	6.378
7	16.843	7.777	7.294	5.831
8	17.263	6.390	6.851	6.046
9	18.385	5.981	6.523	5.461
a	18.469	5.627	6.323	5.897
b	18.370	5.566	2.006	6.338
c	19.302	5.835	7.536	5.723
d	18.391	5.654	6.068	5.610
	(b) Female Students Only			
1	16.430	4.805	5.347	6.998
2	17.763	3.846	6.160	6.108
3	16.127	4.358	6.520	6.246
4	17.585	8.277	5.465	5.253
5	19.409	5.817	7.407	6.283
6	18.721	3.776	5.984	6.048
7	15.364	6.696	6.997	6.040
8	14.908	5.893	5.465	5.778
9	18.385	5.981	6.523	5.461
a	18.469	5.627	6.323	5.897
b	18.370	5.566	2.006	6.338
	(b) Male Students Only			
1	18.678	6.402	6.717	6.540
2	20.185	5.075	4.427	6.549
3	14.893	8.119	6.000	5.933
4	15.577	1.745	6.885	7.792
5	19.308	7.001	6.833	7.146
6	18.060	6.782	6.486	6.739
7	18.260	6.754	7.006	5.593
8	19.075	5.332	6.655	6.307
c	19.302	5.835	7.536	5.723
d	18.391	5.654	6.068	5.610

How adequate is the representation (6.1)? An indirect route is to take the four scores E_i, M_i, Q_i, V_i and subject their correlation matrix to a factor analysis. A more direct approach is to form the estimates as in (6.3) and examine their first two moments and product moments, looking in particular at the covariances (or, if they are easier to interpret, the correlations). These analyses have been done in all colleges, and repeated within mixed-sex colleges for females only and males only. The estimates of $\sqrt{S_1^2 + \frac{1}{4}s^2}$, $\sqrt{S_2^2 + \frac{1}{4}s^2}$, $\sqrt{(S_\Delta^2 + \frac{1}{4}s^2}$, and $\frac{1}{2}s$, are shown in the last four columns of Table 6.1. It is hardly necessary to undertake any formal statistical test to reject a null hypothesis that $\{v_{i2}\}$ and $\{\Delta_i\}$ are components of no significant effect. By using the last column of the Table to give a common estimate of $\frac{1}{4}s^2$ for the other three columns, we can find estimates of S_1, S_2, S_Δ as shown in Table 6.2.

The standard deviations shown in Table 6.2 are certainly of similar order. The median estimate of the error variance $\frac{1}{4}s^2$ for all students is $6.089^2 \approx 37$, which is about the same as for female students only (6.048^2) and a little smaller than for males only (6.307^2 to 6.540^2). The estimate 150 is a convenient figure to use in general calculation for s^2. Recall that below equation (5.5) we estimated the measurement error of $V_i \pm Q_i$ as 18.2^2, so if we regard this as estimating $\text{var}(e_{iV}) + \text{var}(e_{iQ}) = 2s^2$ in the notation of (6.2), then should compare $\frac{1}{2}(18.2^2) = 165.6$ with the estimate 150 as just deduced: the figures are certainly similar. Recall that we have estimated $\frac{1}{4}s^2$ from (6.3d) on the assumption that all terms $e_{i\cdot}$ in (6.2) have a common variance, and that, if anything, we should expect the variances of e_{iV} and e_{iQ} to be larger than the other two variances, which is what we have in fact observed.

Finally, inspect the entries on the means in Table 6.1 (note that the mean for the sum reflects our use of a mean 0.0 in place of 150.0). The mean for a college of the mode of assessment difference ("MADiff") is approximately zero by construction, but when dissected on a gender basis, it is seen to be positive for males and negative for females. It cannot be regarded as a statistical artefact, as Masters and Beswick (= M&B) alleged, whereas it can be regarded, contrary to Masters' advice to the Review Committee that wrote *MATHEF*, as an educational measurement phenomenon, because all that we have done is to classify according to sex the estimates of a measurement contrast Δ_i which by the non-trivial nature of S_2 is shown to be a characteristic of every student in the population and not just a female/male student difference.

TABLE 6.3

Three Gender-linked discrepancies, 1984 ACT data

College	E – ASAT-V	M – ASAT-Q	TE/3.6 – ASAT-T
copc	2.2	2.5	2.5
darc	4.1	1.8	1.5
dckc	6.6	1.7	4.3
ernc	1.6	7.9	2.9
hwkc	1.6	4.0	1.9
narc	4.1	3.7	4.8
phlc	5.5	0.6	3.4
strc	6.3	5.5	4.9
Weighted ave. (1984 scale)	4.3	3.2	3.4
ditto (1985 scale)	7.2	5.3	5.7

Source: Critique to 1986 Review Committee, based on data from Melb. Univ. Centre for Study of Higher Education (Dean McKenzie) and ACT *Year 12 Study*.

Table 6.3 gives results of similar analyses done on 1984 data but weighted by the level (minor, major, major-minor, double major) at which the course score could be included in students' TE scores. The scale represents 0.6 times that used onwards from 1985: the last row repeats the summary data in the 1985 scale, because they can be compared directly with differences between MADiff means under (b) and (c) of Table 6.1 for the eight mixed-sex colleges, yielding $5.5 = 1.828 - (-3.655)$, 2.9, 4.3, 6.9, 5.8, 4.1, 4.3, and 6.8. The global figure for 1984 of 5.7 is quite similar to the average of the 1986 data restricted to English and Mathematics.

Table 6.4

Correlations and Covariances of Sums and Contrasts of Table 6.1 (All Students)

College	df	Sum/QV	Sum/MA	QV/MA	Sum/Err	QV/Err	MA/Err
				(a) Correlations			
1	175	–0.155	–0.080	0.136	–0.119	0.074	0.322
2	169	–0.012	–0.099	0.147	–0.018	0.109	0.471
3	212	0.073	0.103	0.092	0.124	0.042	0.404
4	79	–0.017	–0.052	0.117	0.015	0.064	0.564
5	244	–0.068	–0.128	0.211	–0.104	0.183	0.381
6	178	0.136	–0.113	0.057	0.117	0.243	0.360
7	174	–0.037	–0.110	0.114	–0.061	0.259	0.268
8	112	–0.058	–0.123	0.101	–0.138	0.101	0.243
9	110	0.020	–0.122	–0.183	0.100	0.040	0.083
a	56	0.263	–0.039	–0.111	0.104	–0.006	0.314
b	116	–0.228	–0.129	0.097	–0.021	0.148	0.090
c	72	0.138	–0.061	0.228	–0.062	0.117	0.369
d	120	0.010	–0.100	–0.043	0.041	0.049	0.272
				(b) Covariances			
1		–28.32	–14.37	12.35	–15.32	4.82	20.61
2		–2.12	–16.86	10.70	–2.32	6.00	24.98
3		11.21	15.58	7.57	12.68	2.34	22.15
4		–2.83	–9.02	10.74	1.78	4.03	36.96
5		–13.87	–26.76	21.44	–14.32	12.24	26.12
6		23.97	–20.18	4.70	14.58	13.98	20.99
7		–6.41	–18.31	10.35	–6.34	14.68	14.59
8		–9.33	–20.56	8.12	–15.26	5.37	13.42
9		3.11	–19.90	–12.61	10.47	1.77	3.86
a		41.56	–6.54	–7.82	11.89	–0.29	16.01
b		–37.37	–16.67	5.44	–2.59	7.91	3.79
c		22.71	–11.62	17.63	–7.14	5.47	19.98
d		1.53	–15.89	–2.83	4.42	2.19	12.61

Is this gender-linked discrepancy coupled with the stronger gender-linked difference in the Verbal and Quantitative areas ("QVDiff") as suggested by Masters and Beswick? At face value there is no evidence from Table 6.4 of significant correlation between the two measures Δ_i and v_{i2} because only one of the thirteen sample correlation coefficients is significantly different from 0 (assuming a bivariate normal distribution is appropriate).

Conclusion 6.3. **Analysis via a linear representation gives no evidence of association between the gender-linked discrepancy between course and ASAT scores and the known gender-linked difference in verbal and quantitative skills.**

We deliberately wrote "at face value" because, at least when Sum is one of the component scores, the correlations here entail division by $\sqrt{\text{var(Sum)}}$, and these are about double the other standard deviations (cf. Table 6.1). Using instead the covariances from part (b) of Table 6.4, shows that all these covariances are of the same order.

If in place of a common error variance we suppose that the error terms in equations (6.2) have variances s_E^2, s_M^2, s_V^2, s_Q^2, that are not necessarily all equal, then on the assumption that the main effect variables v_i, v_{i2}, Δ_i are uncorrelated, we should observe for the estimated correlations in

Table 6.4 the quantities

$$
\begin{aligned}
\operatorname{cov}(\mathrm{Sum}, QV) &= \tfrac{1}{16}(s_E^2 - s_M^2 + s_V^2 - s_Q^2), && (6.5a)\\
\operatorname{cov}(\mathrm{Sum}, MA) &= \tfrac{1}{16}(s_E^2 + s_M^2 - s_V^2 - s_Q^2), && (6.5b)\\
\operatorname{cov}(QV, MA) &= \tfrac{1}{16}(s_E^2 - s_M^2 - s_V^2 + s_Q^2), && (6.5c)\\
\operatorname{cov}(\mathrm{Sum}, \mathrm{Err}) &= \tfrac{1}{16}(s_E^2 - s_M^2 - s_V^2 + s_Q^2), && (6.5d)\\
\operatorname{cov}(QV, \mathrm{Err}) &= \tfrac{1}{16}(s_E^2 + s_M^2 - s_V^2 - s_Q^2), && (6.5e)\\
\operatorname{cov}(MA, \mathrm{Err}) &= \tfrac{1}{16}(s_E^2 - s_M^2 + s_V^2 - s_Q^2), && (6.5f)
\end{aligned}
$$

On the basis of what is known about measurement errors in scores, we should expect the relative magnitudes of these error variances to satisfy

$$s_V^2 > s_Q^2 > s_M^2 \quad \text{and} \quad s_V^2 > s_E^2 > s_M^2\,. \tag{6.6}$$

This means that in equations (6.5a) and (6.5f) the quantities should be positive, in (6.5b) and (6.5e) negative, and in (6.5c) and (6.5d) of unknown sign. The second, fourth and sixth columns bear out this prediction, though it does depend on assuming that the substantive variables in (6.5) are uncorrelated; only the terms $\operatorname{cov}(QV, \mathrm{Err})$ indicate any possible contradiction, and these are barely large enough to do so.

It is of some interest to use again the estimates from Chapter 5 to match the correlations of Table 6.4. In the right hand side of (6.5f) we know from Chapter 5 that $\frac{1}{16}(s_V^2 - s_Q^2) \approx (215 - 115)/16 = 12.5$, so if we regard the rest as being about half this, we should have (say) $\operatorname{cov}(MA, \text{Err}) \approx 16$ to 20. From Table 6.1 we use $\operatorname{var}(MA) \approx 83$ and $\operatorname{var}(\text{Err}) \approx 37$, and thus a typical value for $\operatorname{corr}(MA, \text{Err})$ is

$$(16 \text{ to } 20)/\sqrt{37 \times 83} = 0.30 \text{ to } 0.37; \tag{6.7}$$

the median value in the last column in Table 6.4 is 0.322, so we have consistency as regards order of magnitude.

Use of Mode of Assessment Discrepancy Measures

In principle the quantities Δ_i can be used to study groups other than the gender groupings and which may consistently demonstrate a course and ASAT score discrepancy. This is one reason for having discussed the analysis at some length in a report that is concerned with scaling procedures *per se* rather than the sex bias problem which for some time has been treated at a political[17] level rather than technical.

Writing Task and Multiple Choice Verbal Scores

It is a simpler matter to note the further support for the mode of assessment explanation as the origin of the sex bias problem coming from the Writing Task and multiple choice ASAT Verbal sub-scale scores. While these two sets of scores reflect skills which at face value are strongly related, the correlations in Chapter 5 show that the scores $\{Q_i\}$ and $\{V_i\}$ are mutually closer than $\{V_i\}$ and $\{W_i\}$. Part of this explanation may lie in $\{V_i\}$ and $\{W_i\}$ having larger measurement errors than $\{Q_i\}$, but the evidence indicates that it is far from the whole explanation.

[17] See discussion in Chapter 7, especially footnotes 2 and 3.

TABLE 6.5

Within-gender differences and gender-linked discrepancies of Mean Writing Task and ASAT-Verbal scores, 1986 and 1987

	1986 Scores			1987 Scores (approx.)		
College	Differences		Discrepancy	Differences		Discrepancy
	Female	Male		Female	Male	
copc	3.65	–4.04	7.69	0.11	–4.63	4.74
dckc	0.13	–9.77	9.90	–0.61	–7.94	8.55
ernc	–0.52	–9.29	8.77	1.76	–7.79	9.55
hwkc	2.98	–12.53	15.51	1.05	–7.21	8.26
narc	–2.32	–7.54	5.22	0.17	–4.05	4.22
phlc	6.00	–11.59	17.59	0.81	–10.15	10.96
strc	2.09	–8.01	10.10	6.17	–1.00	7.17
darc	7.91	0.11	7.80	9.42	4.30	5.12
ccec	9.88			1.00		
merc	11.57			6.24		
stcc	12.79			6.89		
edmc		1.82			4.56	
marc		0.31			2.78	
All Govt.	1.72	–8.97	10.69	1.38	–6.27	7.65
All non-G.	10.54	0.75	9.79	5.61	3.85	1.76
System	4.73	–5.86	10.59	2.80	–3.39	6.19

Source: Year 12 Studies and the 1986 datum (inferred from rescaling of $\frac{1}{2}(V + W)$ to ACT verbal score) that corr(ASAT-V, Writing Task) = 0.4943.

Table 6.5 lists the differences between mean Writing Task and ASAT-V scores for the ASAT score populations in each college, as can be found from Tables 8 of the ACT *Year 12 Study*. These data show unequivocally for 1986, and less noticeably for 1987, that across the system, once there is control of the schooling experience, females tend to have better scores on the essay-writing assessment in the verbal area and males better in the multiple choice assessment. Again, note that the data are based on just two sets of papers – though two are better than one as for most of this report. It is particularly significant that the discrepancy is consistent with other data.

Another noticeable effect is that for both females and males, there is a difference in the mean scores for students within the Government sector and those outside, with the latter having better scores on the Writing Task and students in the Government sector having relatively better scores in the multiple choice test.

CONCLUSION 6.4. **Test results for both 1986 and 1987 showed systematic differences in assessment by essay-writing and by multiple choice tests, with respect to both sex and school type.**

Chapter 7

Bias Problems with Reference Scales

Our discussion to this point has mostly been about the construction of scale as distinct from location parameters, i.e., about "getting the right standard deviations" of scores, because it is in this respect that various scaling *procedures* differ most markedly. The variety of scale parameters from the different procedures has evolved out of presumption rather than proof. The problems and techniques are necessarily more involved than the matter to which we now turn briefly, namely, to potential problems of bias between reference scale and school-based scores: they are real problems in the ACT and Queensland.

We saw in Conclusion 6.1 that discrepancies in assessment between school-based and ASAT multiple choice test scores exist for individual students, irrespective of the educational "dimension" being Quantitative or Verbal. These discrepancies are not quite equally distributed between females and males, so they are gender-linked. We exhibited other gender-linked differences and discrepancies in Chapter 6 also.

The reason such gender-linked discrepancies matter is that in both the ACT and Queensland, the system authorities wish to assert that the multiple choice test based ASAT scores can be used as a common scale on which the scores of entire schools are placed "fairly" by using the ASAT scores for schools as a whole. No matter how it is done, these ASAT scores effectively provide the means and standard deviations of schools' Tertiary Entrance scores. The claim for the scores of the schools being "fairly" placed on the scale is made irrespective of the school being mixed- or single-sex, yet as seen in a literature that goes back decades (Daley, 1986a), multiple choice scores exhibit a consistent gender-linked discrepancy relative to school-based assessments. Therefore, the use of multiple choice tests purportedly to place school-based assessments on a common scale when the gender mix of the groups varies all the way from 100% male to 100% female, introduces an effect from the gender-linked discrepancy, and leads to bias in the resulting TE scores as reflecting school-based assessment as claimed.

To see that a bias and substantial correlation can coexist, suppose given two samples of the same size from a bivariate distribution with correlation coefficient r. Increase one of the variables by a fraction b of its standard deviation in one of the samples, and decrease it the same amount in the other sample. In standardized units, this gives a discrepancy between the two variables in the two samples of size $2b$, while elementary algebra shows that the correlation coefficient changes to

$$(r - b^2)/(1 + b^2) \approx r - (1 + r)b^2, \tag{7.1}$$

which, for example, for $r = 0.7$ and $b = 0.15$ equals 0.66. One contributing factor to the slightly lower correlations between ASAT and TE scores consistently observed in the ACT in mixed- rather than single-sex colleges, is the gender-linked discrepancy between the two scores (typically, b is in the range 0.1 to 0.2).

The Original Complaint

A bias effect showed itself in the ACT in the first year of TE scores in 1977. Such shifts as seen in the Girls' schools column of Table 7.1[18] prompted an immediate query that started the gender-linked discrepancy discussion in the ACT. In 1976, students from the ACT high schools and colleges who twelve months later would have received ACT TE scores, had HSC aggregate scores uniformly distributed over the range of NSW HSC aggregate scores. The changes between the 1976 HSC aggregate and 1977 TE score distributions are given for the two groups of students in single-sex ACT colleges, along with the discrepancy. The data from the female colleges alone show a significant shift (use a Kolmogorov–Smirnov test). Taken together, the female and male college data show that there was a gender-linked shift. Treat the 1976 HSC scores as though they were the 1977 TE scores from the mixed-sex colleges; then Discrepancy in Table 7.1 corresponds to a gender-linked discrepancy as computed in Tables 6.3–5 of from 9.5 to 12.5 points on the 1985 scale (cf. 5.7 in Table 6.3, 10.6 and 6.2 in Table 6.5).

Table 7.1

Shifts in Proportions of ACT Single-Sex College Students in Ranking by 1976 NSW HSC Aggregates and 1977 ACT TE Scores

Top proportion	Girls' schools	Boys' schools	Discrepancy
10%	– 5%	+ 11%	16%
20%	– 6%	+ 9%	15%
40%	– 12%	+ 10%	22%
60%	– 15%	+ 4%	19%
80%	– 14%	0%	14%
Numbers: 1976	281	220	
1977	291	248	

Sources: 1976 data from hardcopy output of NSW results made available to ACT in preparation for 1977 change. 1977 data from ACT Schools Accrediting Agency report on 1977 TE scores (no *Year 12 Study* was produced for 1977 results).

The assertion (*MATHEF*, §7.1) that

> "... [because] the major source of [the sex-bias problem] is the multidimensionality of the scores ... that are combined to form [an aggregate score, the problem is] not unique to the Australian Capital Territory or unique to a system operating without external examinations" ,

is little more than wishful thinking because it is contrary to the "two trends that suggested the possibility of bias" (§1.2 of *MATHEF*). We now also see that it is contradicted by the evidence in Table 7.1. Fortunately, the Review Committee gave it as "a view formed on balance on the basis of conflicting evidence". Regrettably, this basic evidence was never produced to the Committee, nor sought by them nor any of their predecessors, save perhaps Morgan (1979). I acquired data to compile it only when *MATHEF* was being received, so it was never discussed by the Committee.

[18] The data in Table 7.1 were first assembled *c.* June 1986, after the report *MATHEF* had been compiled. In July 1986 they were tabled at a meeting of the Accrediting Agency which has never referred to them. They were included in a conference paper (Daley, 1986b) appended to Daley (1987b).

Some Average Achievement and Test Scores

Table 7.2 summarizes various mean scores for the thirteen colleges in the ACT where scores are scaled by a statistical scaling procedure. They are given for all female, all male, and all students at each college. The scores concerned are

TEACT ACT TE score as issued in 1986;

TEMM Quasi TE score using Method of Moment estimation in an Other Course Score scaling procedure;

ASAT Total ASAT score as used in 1986 (roughly, 50 : 35 : 15 $Q : V : W$ mixture);

v_i The scaling criterion variable used in compiling TEMM;

AQU, AVB, AWR ASAT sub-scale and Writing Task scores;

OPTAT Optimal mixture of AQU, AVB, AWR as in Unif. Opt. in Tables 4.2–4.

The tabulated quantities have a system-wide mean of 0.0 (rather than 150.0) for ASAT and sub-scale scores, and the TE scores are in fact divided by 3.6 (= number of course scores in best 3.6 aggregate).

The scaling procedure ensures that the mean ASAT and v_i scores are approximately the same. TEACT exceeds ASAT and TEMM exceeds v_i by a similar amount for all colleges because of the "best 3.6 course scores" selection effect. *NBNBNB: Check this para. with original*

OPTAT is similar in structure to ASAT except that it uses more AWR and less AVB. The difference (ASAT) – (OPTAT) thus reflects a college's standing in these two scores: not much change for the Government mixed-sex colleges nor single-sex male colleges, but increases for single-sex female colleges. Dissected by sex, there are decreases for males at Government colleges. Had OPTAT been used in place of ASAT scores in producing TEMM scores, the gender-linked discrepancy would have been reduced from a mean of about 6.4/25.0 = 0.25 standardized points to about 4.4/25.0 = 0.18 standardized points. TEMM scores are generally decreased from TEACT scores because the latter were calculated using more than one reference scale score, thereby introducing selection bias effects, being more marked for males than females as sub-scale scores were used more often for males (Mathematics, Physics, Chemistry) than for females.

The assumption at A 5 that "ASAT scores are sufficiently strongly correlated with measures of general relative academic achievement to make it feasible to rescale course scores to a 'common scale' valid for comparing student achievement across all schools within the system", can only be as valid as the stated or unstated assumptions which it entails. The most critical of these is that ASAT and school-based scores measure a similar variable, whether labelled achievement or developed ability, in the primary "dimension" in which schools differ, *subject only to measurement (and model-fit) error.* What is the extent of this error? And how do school-populations and their sets of course scores differ?

To answer the second question first, school-populations differ (i) by their ASAT score distributions; (ii) by gender-mix; (iii) by both numbers and proportions of students seeking a TE score; and (iv) by the length of time the students have attended the school or college. In terms of how these differences may impinge on the ASAT/TE score relationship, (i) reflects the primary dimension of assumption A 5; the gender-relationship (ii) is the one most explored; I have seen some evidence re (iv) which is consistent with students entering an existing group having on average marginally lower TE scores in relation to their ASAT scores than for the rest of the group they are entering (but this is a smaller effect than the gender-linked discrepancy); I have not been able to discern a systematic relation reflecting (iii).

Table 7.2

Averages of general achievement and ASAT sub-scale measures

Run #	TEACT	TEMM	ASAT	v_i	AQU	AVB	AWR	OPTAT
(a) All Students								
620	7.14	6.45	1.70	1.94	-1.53	0.92	11.02	4.33
621	0.31	-2.11	-5.36	-6.02	-2.83	-5.71	-6.07	-5.78
622	5.44	5.06	0.70	0.93	0.51	-0.42	3.13	1.63
623	5.03	2.87	-0.89	-0.99	0.50	-0.49	-4.86	-2.03
624	6.63	5.00	0.80	1.25	4.37	-3.06	-1.31	1.25
625	3.86	0.72	-1.96	-2.27	0.33	-1.80	-6.84	-3.35
626	7.81	6.81	1.61	1.91	1.35	2.59	-1.64	0.66
627	7.72	6.29	1.51	1.84	6.11	-3.01	-2.76	1.63
628	-2.42	-3.57	-6.67	-7.67	-11.87	-3.52	8.07	-4.14
629	8.13	6.74	0.56	1.28	-0.04	2.18	-2.10	-0.47
630	10.41	9.24	4.74	5.53	5.17	3.46	1.44	4.51
631	5.97	5.42	0.18	-0.14	-8.24	3.29	16.95	3.65
632	5.65	3.13	-1.44	-1.64	-0.65	-0.93	-3.36	-2.13
(b) Female Students only								
621	-2.43	-3.49	-10.95	-6.97	-11.83	-7.54	-4.65	-10.86
622	2.26	2.10	-3.19	-1.38	-7.76	-0.63	7.35	-1.37
623	6.41	4.71	-1.17	1.24	-2.93	0.67	0.83	-1.19
625	5.07	3.72	-6.75	0.29	-7.97	-3.34	-3.99	-7.30
626	6.90	7.04	-1.39	2.24	-5.87	1.98	5.12	-0.66
629	8.39	7.53	0.28	2.58	-3.74	4.34	2.05	-0.25
630	9.90	9.33	2.21	5.60	-2.25	3.49	9.70	3.92
632	4.05	2.89	-7.18	-1.70	-8.91	-3.98	-2.03	-7.11
620	7.14	6.45	1.70	1.94	-1.53	0.92	11.02	4.33
628	-2.42	-3.57	-6.67	-7.67	-11.87	-3.52	8.07	-4.14
631	5.97	5.42	0.18	-0.14	-8.24	3.29	16.95	3.65
(c) Male Students only								
621	3.58	-0.47	1.33	-4.88	7.93	-3.52	-7.77	0.30
622	7.89	7.34	3.71	2.70	6.86	-0.25	-0.11	3.93
623	3.41	0.70	-0.55	-3.61	4.54	-1.86	-11.55	-3.03
625	2.62	-2.34	2.93	-4.89	8.82	-0.22	-9.75	0.69
626	8.82	6.56	5.00	1.53	9.49	3.28	-9.27	2.14
629	7.85	5.86	0.87	-0.17	4.08	-0.23	-6.73	-0.71
630	11.00	9.14	7.70	5.44	13.83	3.42	-8.21	5.21
632	7.68	3.44	5.89	-1.58	9.89	2.96	-5.06	4.21
624	6.63	5.00	0.80	1.25	4.37	-3.06	-1.31	1.25
627	7.72	6.29	1.51	1.84	6.11	-3.01	-2.76	1.63

Concerning the first question, the correlation between ASAT and school-based measures of general relative achievement is generally around 0.70, implying that a school's ASAT scores explain about 50% of the variation in its set of achievement measures, whereas the signal to noise ratio of the latter is typically 1.6 to 3.6 so that the general measures v_i explain from 60% to 78% of the variation (recall also Conclusion 4.7). Thus, there is ample scope for systematic effects to exist between ASAT and school-based achievement measures: this is precisely what we showed in Chapter 6 by demonstrating the existence of a systematic difference in mode of assess-

Figure 7.1

Masters and Beswick's Groups M and N data: Regression to the Mean

(from p.32 of *MATHEF*)

ments for the student body as a whole, not merely as a difference between the groups of females and males within mixed-sex colleges. And as best we could tell (cf. Table 6.4), this difference is not associated with the stronger difference in the Verbal and Quantitative areas ("QVDiff") as Masters & Beswick (= M&B) suggested.[19]

Masters and Beswick's Evidence

The only substantive evidence that M&B offered to support their claim for a link between courses studied and the gender-linked discrepancy between ASAT and course scores, involved their analyses of such data in what they called Groups M and N (roughly speaking, students whose TE scores have a substantial component of Mathematics and science scores, or Not). They gave their results in graphs like Figures 5.3 and 5.4 of *MATHEF*, reproduced here. The dominant feature in them is the regression to the mean effect which they properly called a predictable bias, i.e., a statistical artefact, and not a bias *per se* (§§3.29–32 of their report). This implies that any discrepancy of the two measures is well summarized by comparing the means of the groups concerned as in Table 7.3. Two features in the table are obvious:

(a) All gender-linked discrepancies are positive in spite of an implied tendency towards "unidimensionality" of course scores within each of the groups M and N.

(b) For Group M, there is a noticeable discrepancy between the shifts in ASAT and TE scores from 1984 to 1985: this underlines the variable dependency on the ASAT paper of both gender-linked differences of ASAT scores and the gender-linked discrepancy between them and TE scores (see Conclusion 5.2).

[19] *MATHEF* (§1.19) offers the following reasons for the Accrediting Agency's persistent refusal to heed the advice of its Technical Advisory Committee to remove the gender-linked bias affecting TE scores from the single-sex schools in particular: "[There] were those who were reluctant to believe that sex *per se* could be the cause of the observed inconsistency between teacher assessments and aptitude test scores. They claimed that no policy change should be introduced until further research had been undertaken to clarify the bases of the sex differences. They suggested that the sex differences might be due to (i) the pattern of course enrolments of the students, particularly in humanities and mathematics/science courses; or (ii) a bias in teacher assessments in favour of female students rather than a bias in aptitude test scores in favour of males."

Four comments on these statements are relevant. (1) The ACT system asserts that teacher-based assessments provide the standard for certification. Granted that female and male children do not have identical upbringing, and granted that asking them to execute certain tasks (ASAT paper, or write an essay) may evoke different behaviour, the presumption that sex should have no effect on assessments can only be viewed as convenient unless shown otherwise. Morgan (1979) was the first to do so using ACT data. (2) There is no policy change involved in producing TE scores that reflect teacher-based assessments consistently for all students. Rather, it is illogical for the Accrediting Agency to continue to pretend that TE scores have been and are being issued as equally fair to all students as far as is possible when they fail to reflect teacher-based assessment in an unbiased manner. (3) Masters and Beswick were commissioned to undertake research that would investigate the causes of observed sex differences, and the only explanation that they produced was that these differences arose as a statistical artefact in the production of an aggregate. The evidence offered in support (their analyses involving "Groups M and N") was fundamentally flawed: a valid statistical analysis shows no such support (see Table 7.3). Further, a statistical artefact should be capable of demonstration on the basis of extreme assumptions that if anything accentuate the cause; Masters and Beswick offered no such demonstration, and I have tried and failed to provide such. This latter failure is not surprising, because all the evidence points to the problem being a measurement problem, not a statistical artefact. (4) If there is sex-linked bias in teacher assessments then a basic assumption (A 2) of the entire assessment system is incorrect. To my knowledge no evidence to test the suggestion has been sought, nor am I aware of any evidence from outside that supports it.

Table 7.3

Mean TE and ASAT Scores in Mixed-sex Colleges for Masters and Beswick's Groups M and N.

		Group M		Group N	
		1984	1985	1984	1985
Females	Numbers	372	406	413	373
	3.6× ASAT	253.23	243.11	208.30	204.91
	TES	264.15	261.15	227.89	224.29
	3.6× ASAT-V	252.7	N/A	224.3	N/A
	3.6× ASAT-Q	249.5	N/A	198.0	N/A
Males	Numbers	532	514	236	200
	3.6× ASAT	250.60	254.87	212.64	204.05
	TES	255.68	256.63	213.79	205.48
	3.6× ASAT-V	236.9	N/A	216.7	N/A
	3.6× ASAT-Q	261.0	N/A	213.1	N/A
Gender-linked Discrepancy		5.85	16.75	18.44	17.95
do. relative to ASAT-Q		20.0	N/A		
do. relative to ASAT-V				6.7	N/A
do. after approx. calibr'n		– 5.1	– 0.5	7.5	0.7

Source: Copy from Dr. G. Masters in July 1986 of data processed for Masters and Beswick, and Table 6b in their report. I thank Dr. Masters for the detailed data.

The discrepancy is about the same for both groups in 1985, indicating no apparent association with the "dimensions" represented by Groups M and N. Yet, it is the data of 1984 that are closer to what would be expected *a priori* on the basis of the known gender-linked differences given by QVDiff. To see this, argue as follows. Make the "dimension" of the ASAT reference scores line up more closely with the presumed dominant "educational domains" of the components of the TE scores of the two groups by replacing ASAT-T scores for Groups M and N by $\{Q_i\}$ and $\{V_i\}$ respectively. Small increases in total TE scores will result, accompanied by somewhat larger shifts of the female and male ASAT scores: fair-sized increases for females in Group N and males in Group M, and moderate to smaller decreases for females and males in the other groups respectively.

Now check these predictions against data (and, for 1984 data I can only use the existing TE scores, but they will be little changed, as shown by analyses at the ACT Accrediting Agency in 1984 in which both adjusted and sub-scale ASAT scores were used in place of ASAT scores). Gender-linked discrepancies remain, though the relative standing of Groups M and N has interchanged because in reality we have used a reference measure favouring males in Group M and females in Group N, and have probably overcompensated.

On the other hand, removing a common gender-linked discrepancy effect as under calibration leads to a position closer to no discrepancy in either group for 1984 data, and none for 1985 data. The approximate position reached in the last line of Table 7.3 is about the closest position to unbiasedness that can be attained using only the information available in 1984.

Of our observations, none supports M&B's explanation of the gender-linked discrepancy as being a statistical artefact attributable to different patterns of course selection, while others refute it by exhibiting the discrepancy as a mode-of-assessment phenomenon or by contradictory data as in Table 7.1. Had M&B used a statistical model to describe the scores, they may then have been in a position to deduce algebraically how the alleged statistical artefact arises, because it should be feasible to exhibit any such artefact in algebraic terms. I have tried to produce the claimed artefact via the route alluded to in M&B, or any other way. The only "bias as a statistical artefact" I could concoct relied on a differential correlation coefficient effect that was discussed at an Accrediting Agency meeting in mid-1985 (cf. McGaw, 1987). It was noted then that by making extreme assumptions, about 10% of the bias as computed from data can be explained. The effect comes from using a sub-optimal scaling procedure, rather than the aggregating operation itself, because it vanishes on using Method-of-Moment estimation in an Other Course Score procedure which is self-consistent.

Instead of such algebraic calculation, M&B resorted to imprecise argument based on contingencies whose effects were not measured, then argued that the observed gender-linked discrepancies were consequences of the unmeasured effects in an undefined model, and concluded that this supported their hypothesis (cf. *MATHEF*, §§5.29–30).

It is far from clear why the ACT Schools Accrediting Agency has allowed[20] the bias in TE scores to persist so long.

Conclusion 7.1. **The gender-linked discrepancies between ASAT and course scores that result in gender-linked biases in Tertiary Entrance scores reflect different processes for measuring educational properties. The discrepancies are not a statistical artefact of the process of aggregation**.

Gender-linked Biases and Other Course Score Scaling

For the record, we construct from Table 7.2 the entries in Table 7.4 of the gender-linked discrepancies between TEACT and ASAT, TEMM and ASAT, and TEMM and OPTAT. A large part of the difference between TEACT and TEMM against ASAT comes from the use in TEACT of sub-scale scores as reference mea- sures, whereas in TEMM there is only one reference measure used and the scaling criterion variable is defined consistently via Other Course Scores. The change in moving from ASAT to OPTAT is a consequence of changing the relative weighting of Writing Task (increased) and ASAT-V (decreased). In no sense would one suggest, on this single data set, that the gender-linked discrepancy is made noticeably different on account of using an Other Course Score scaling procedure.

How can the bias be removed? The surest way is by statistical calibration as first canvassed in 1984. The Accrediting Agency's reasons for rejecting the operation then are now even less valid, though its political unpalatability may well be unchanged. Technically, it is the preferred method, not least because it ensures that the reference scale scores correlate optimally with the school-based measure of general achievement (it follows from (7.1) that reference scale scores that incorporate a bias correlate better on removing the bias).

[20] While one can only speculate on motives, resistance to the principle of producing fair TE scores by whatever means available, may be based amongst other reasons on any or all of (i) boys' schools not wishing to lose an advantage; (ii) the inability to justify an *arbitrary* statistical adjustment on the basis of gender in 1983; (iii) an inability to distinguish between a gender-linked *difference* in a single set of scores and a gender-linked *discrepancy* between two sets of scores that are positively correlated and supposedly measure the same variable subject only to measurement error; (iv) unwillingness to disturb the *status quo* because to do so may foster pressure for the return of an external examination system; (v) unwillingness to declare that the Scholastic Aptitude Test scores "objectively" determined by an independent outside body, the Australian Council for Educational Research, can be other than 100% adequate for the purpose for which they are used; (vi) unwillingness of the Educational Measurement Testing industry to acknowledge that assessments based on multiple choice tests can be biased on the basis of culture or gender relative to another method of assessing the ability or achievement of students, and that such other assessments may have equal or greater validity.

In 1986 the Accrediting Agency showed that it saw the matter as political and not technical by sacking and not replacing its Technical Advisory Committee, so losing years of accumulated knowledge.

TABLE 7.4

Gender-linked Discrepancies (= Bias Measures) between Two TE-like Scores, and ASAT and an ASAT-like Scores

Run #	TEACT/ASAT	TEMM/ASAT	TEMM/OPTAT
621	6.27	9.26	8.14
622	1.27	1.66	0.06
623	3.62	4.63	2.17
625	12.13	15.73	14.04
626	4.47	6.87	3.28
629	1.13	2.26	1.21
630	4.39	5.68	1.48
632	9.44	12.52	10.77

A second route, consistent with the evidence of Chapter 6 and much more besides (Daley, 1986a), is by raising the extent of traditional system-wide assessment: in other words, relying much more on some external examinations. This is unpopular with the power-base of the ACT Schools Authority, being contrary to the principles espoused in the mid-seventies when the ACT moved away from this time-honoured[21] assessment system (cf. P 3). There is also some practical difficulty in having external examinations in only a portion of the curriculum, as distorted results may come from different colleges placing more or less emphasis on teaching externally *v.* internally examinable material.

A third possibility which has not been investigated[22] is the use of an apparently bias-free scale like ASAT-*T* in Mathematics scores (or, Mathematics and exact Sciences, with or without Computing; cf. M&B), followed by using an Other Course Score scaling method to produce scores in all courses much as has been done in Chapter 4 where the scale coefficient of the external reference scale scores was fixed. To adopt this route, one would first use the approach of Chapter 4 to find an optimal pre-dictor in terms of $(Q_i \; V_i \; W_i)$ for (say) Mathematics scores, scale these by a two-moment bivariate equating method, assuming that the error terms have similar variances, and finally execute a Method-of-Moment estimation procedure on the various college data sets treating the scaled Mathematics scores as fixed (so $\beta_{\mathrm{Math}} = 1$). As a scaling procedure this is far from as efficient as the one developed earlier.

A fourth possibility noted in Chapter 1 is to increase the weighting of Writing Task scores in the "Total ASAT" score. This route does nothing to control the annual variability in gender-linked differences (Conclusion 5.2), and it leads to a reference scale score that is sub-optimal in the sense of Conclusion 4.2.

CONCLUSION 7.2. **ASAT scores, with or without Writing Task scores, are not sufficiently positively correlated with school-based assessments to justify rescaling course scores or general achievement measures without considering the necessity for their calibration to remove the gender-linked discrepancy between them and course scores. Calibrated scores have higher correlations with school-based scores. Other action to check on outlier scores may marginally affect the discrepancy measure.**

[21] The Chinese of *c.* 2000 B.C. are usually credited with introducing state-run public examinations, for entry to the civil service.

[22] A disturbing feature of some recommendations in *MATHEF* was that their likely effects could be studied and predicted using information from both inside and outside the ACT, but even the potential for such appears not to have been recognized. Some predictions that were given in Daley (1986b) are consistent with subsequent observations.

Chapter 8

Imprecision in Calculating Tertiary Entrance Scores

Scope for Jackknife Studies

Jackknife techniques can be used to study the uncertainty in a student's TE score arising from the pooled effect of the other students with whom any given student's TE score is constructed. As well as providing quantitative data concerning this aspect of the precision of TE scores, these studies reinforce previous statements identifying ASAT scores as a more substantial source of imprecision than the school-based assessments. Theoretical work alluded to in Daley (1985, note 5 on p.245) is confirmed by the empirical analyses below. For details on resampling techniques including the jackknife, see e.g. Efron (1982) or Efron & Tibshirani (1986).

Under any Other Course Score scaling procedure, for any given set of parameters (e.g. large and small group sizes, ASAT weight, adjusting for outlier scores), we can compute the effect on TE scores of using estimates $(a_j,\, b_j)$ of the scaling parameters,

(i) for a fixed given set of ASAT scores, by resampling the course scores;

(ii) for a fixed given set of course scores, by resampling the ASAT scores;

(iii) by resampling both sets of scores.

Having found $(a_j,\, b_j)$ for a particular resampled set, TE scores can be found for each student. Under the ACT's or Queensland's existing ASAT scaling procedure, a resampling process analogous to (iii) can be done; with rather more difficulty, an analogue of (i) could also be devised.

Suppose there are N students in the population. A jackknife estimate of the (error) variance of the TE score TE_i of student i based on the N subsets of $N-1$ students consisting of all except the j^{th} student ($j = 1, \ldots, N$), is given by

$$(s^2_{\text{JNF}})_i \equiv \sum_{j=1}^{N}(TE_i - TE_{i,\backslash j})^2 - \frac{1}{N}\sum_{j=1}^{N}(TE_i - TE_{i,\backslash j})]^2 \tag{8.1}$$

where $TE_{i,\backslash j}$ denotes the TE score for student i when scores for the j^{th} student are omitted in the calculation of the scaling parameters.

The one-factor model representation of course and ASAT scores leads directly to the representation for an individual's TE score, namely

$$TE_i = 3.6 v_i + \tau + e_{iT}\,. \tag{8.2}$$

Here, τ represents the selection effect ("best 3.6 scores") making a student's score exceed the one-factor measure v_i (see e.g. Daley, 1985), and e_{iT} denotes model-fit and measurement error. In both theory and practice, the imprecision in TE_i depends on the standardized relative general achievement/ability measure

$$\zeta_i \equiv [v_i - \text{ave}(v_i)]/s.d.(v_i), \tag{8.3}$$

(i.e., a "z-score"). Specifically, theory predicts and empirical studies confirm that

$$(s^2_{\text{JNF}})_i/(1 + \zeta_i^2) \tag{8.4}$$

TABLE 8.1

Jackknife estimates of "imprecision" in a student's TE score due to "errors" in other students' scores in the same scaling group(s)

Run ID #'s		SD TE Scores		Cse. Scs.	ASAT Scs.	Both Scs.	ASAT Scal.
		OCS Scal.	ASAT Scal.				
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
540	501	79.58	82.51	18.2	61.6	76.1	105.1
541	503	79.64	82.52	16.2	58.3	71.5	134.6
542	500	63.28	68.18	17.3	47.0	58.6	105.3
543	512	72.08	68.85	13.5	23.5	34.3	62.9
544	505	74.58	75.64	13.9	33.2	44.6	61.1
545	507	79.84	82.02	16.2	22.5	35.6	46.1
546	504	76.61	79.57	13.5	28.7	39.8	47.4
547	509	72.42	75.78	10.8	16.4	25.3	54.8
548	508	77.07	74.02	17.0	33.3	46.2	81.6
549	506	81.08	79.73	12.9	23.8	35.1	50.9
550	502	71.55	73.72	18.1	22.2	36.9	51.8
551	511	90.54	88.05	26.9	26.0	48.7	94.4
552	510	83.94	89.60	17.4	19.3	34.2	59.0

Notes: The Run ID numbers in (1) refer to columns (3) and (5)–(7), and numbers in (2) to (4) and (8).

The TE Score standard deviations in columns (3) and (4) refer to Other Course Score scaling with Method of Moment estimators for (3), and 1985-style ASAT scaling for (4).

The jackknife variance estimates in columns (5)–(7) are the mean square statistics as at (8.5) corresponding to the resampling schemes as at (i)–(iii) respectively. The estimate in (8) is the analogous quantity from 1985-style ASAT scaling.

is much less variable than the unstandardized quantities in the numerators here. As an overall estimate of the scaling error, we use the mean square statistic, for college k,

$$(S^2_{\mathrm{JNF}})_k \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{(s^2_{\mathrm{JNF}})_i}{1+\zeta_i^2}\,. \tag{8.5}$$

Interpretation

We can interpret these mean square statistics at (8.5) as follows. Consider first an "ideal" student Pat with about an average TE score, so that $\zeta_{\mathrm{Pat}} \approx 0.0$. Suppose Pat attends the college from whose data the statistics in the first line of Table 8.1 have been compiled via Method of Moment estimation in an Other Course Score scaling procedure. Then viewing students' course scores as being subject to "error" around their "true score" values, the uncertainty in Pat's TE score relative to other students in the same college can be described by a random variable with zero mean and standard deviation $\sqrt{18.2} \approx 4.3$. If the students had received exactly the same course scores from school but a different ASAT paper used, the uncertainty in Pat's score is given by a random variable with standard deviation $\sqrt{61.6} \approx 7.9$. When we regard both ASAT and course scores as being subject to random variation, the uncertainty in Pat's TE score relative to other students at other colleges can be described by a random variable with standard deviation $\sqrt{76.1} \approx 8.7$. Had the 1985 ASAT-style scaling procedure been followed (though, using the 1986 ASAT-T scores incorporating Writing Task scores), the last figure would have been higher still at $\sqrt{105.1} \approx 10.2$.

TABLE 8.2

Jackknife estimates of "imprecision" under different variants of Other Course Score scaling procedure with Method of Moment Estimation

Run ID #		Outly. %	Av.TE Sc.	SD TE Sc.	"Imprecision" variance estimates		
	ASAT Wt.				(i)	(ii)	(iii)
540	0	0.00%	–5.174	68.902	15.92	99.28	113.49
		2.23%	–1.103	76.151	16.08	54.22	68.38
	1.5	0.00%	–1.966	79.583	18.16	61.63	76.06
		2.57%	1.752	85.284	17.13	39.08	53.53
541	0	0.00%	13.853	65.838	12.83	84.36	98.55
		4.52%	13.829	69.069	14.52	170.42	184.79
	1.5	0.00%	15.774	79.638	16.18	58.25	71.50
		6.02%	12.027	83.111	17.86	90.49	103.38
542	0	0.00%	2.377	47.306	14.11	72.85	86.69
		2.65%	1.788	47.944	14.44	120.49	137.68
	1.5	0.00%	4.661	63.281	17.29	47.04	58.62
		3.69%	2.027	62.967	18.52	68.02	81.13
543	0	0.00%	–6.996	57.232	12.86	37.16	48.37
		4.41%	–5.649	62.613	12.52	29.62	42.18
	1.5	0.00%	–4.441	72.083	13.49	23.55	34.30
		5.32%	–3.359	74.610	13.78	25.22	37.21
544	0	0.00%	18.159	62.626	10.70	45.23	55.85
		3.96%	19.230	64.309	10.65	59.37	70.21
	1.5	0.00%	20.702	74.578	13.93	33.23	44.58
		4.68%	21.321	74.715	13.53	43.78	55.22
545	0	0.00%	15.486	66.375	11.53	29.31	40.45
		5.06%	17.841	67.517	11.85	33.82	45.20
	1.5	0.00%	18.076	79.840	16.19	22.51	35.55
		5.10%	19.436	78.063	15.27	25.67	38.01
546	0	0.00%	18.823	65.392	10.41	38.26	48.55
		1.60%	19.351	65.130	10.49	53.77	64.56
	1.5	0.00%	21.086	76.614	13.50	28.68	39.77
		1.96%	20.492	76.945	13.89	35.08	46.66
547	0	0.00%	3.685	56.396	9.20	26.84	35.98
		5.08%	7.057	57.650	9.32	35.24	44.91
	1.5	0.00%	5.470	72.421	10.75	16.44	25.33
		5.80%	5.624	73.564	10.61	21.42	30.62
548	0	0.00%	4.689	9.012	15.98	55.67	70.19
		6.00%	9.810	1.870	15.98	52.54	67.39
	1.5	0.00%	8.139	7.068	17.02	33.34	46.21
		.29%	11.318	73.631	16.58	38.29	51.61
549	0	0.00%	12.321	72.744	11.94	27.74	39.82
		2.09%	14.001	73.090	12.07	28.42	40.85
	1.5	0.00%	13.049	81.083	12.90	23.82	35.06
		2.16%	14.186	80.287	13.21	26.10	37.71
550	0	0.00%	28.180	55.625	15.32	31.25	46.51
		5.32%	32.737	54.263	15.59	38.01	53.57
	1.5	0.00%	31.682	71.549	18.07	22.21	36.87
		5.26%	34.768	68.349	18.75	24.83	40.59

(Continued on next page)

TABLE 8.2 (cont.)

551	0	0.00%	22.575	73.531	24.65	39.55	63.06
		7.30%	30.811	69.465	25.54	74.86	98.04
	1.5	0.00%	28.147	90.535	26.88	25.97	48.67
		8.11%	32.834	89.880	27.24	46.13	69.26
552	0	0.00%	19.299	66.858	14.99	28.92	44.22
		7.38%	23.460	71.031	15.10	39.56	55.42
	1.5	0.00%	22.446	83.937	17.38	19.32	34.23
		8.60%	24.771	85.419	16.78	27.19	42.09

Note: The entries under items (i)–(iii) have been rescaled from the jackknife estimates as computed by making all figures have the same SD for the TE score as in the third line of entries for each run.

The entries in column (7) are approximately equal to the sum of the entries in columns (5) and (6). Thus, the two sources of imprecision just considered of uncertainties in course and ASAT scores, when measured by variances, are additive, as we should expect if they and their effects on TE scores were independent.

The TE score standard deviations that result from 1985-style ASAT scaling are about the same or marginally higher than the standard deviations that the Method-of-Moment scaling procedure gives when, as in this table, ASAT scores are weighted like 1.5 courses in computing the measures v_i used in scaling. The comparison between the imprecisions from the two procedures is only reasonable because these TE score standard deviations are approximately the same.

If Pat were a student seeking admission to a course with a higher TE score (e.g. Law) the imprecision would be at least twice the size of the estimates given above. If instead Pat were a student under the NSW HSC system, these estimated standard deviations would be reduced to less than 10% of their present values, and only the entries in column (i) would matter. This last observation helps emphasize that *any* scaling procedure incorporates an element of imprecision; the ideal is to make it as small as possible consistent with the principles that underlie the construction of a TE score.

Table 8.2 lists similar quantities to columns (5)–(7), for the same ASAT-T weighting as in Table 8.1 in the third and fourth lines for each college, for zero ASAT-T weighting in the first two lines, and with adjustments for outliers being made in lines 2 and 4. The estimates have been rescaled by making the TE scores have the same standard deviation as in the third line, rather than their actual values as shown. Such standardization tends to reduce the effect of course score "error" variation as a result of the weighting of ASAT scores. The reduction in the ASAT component of variability (comparison of entries in column (ii) for similar outlier % between different ASAT weights, namely, as a course score weight of either 0 or 1.5), may be attributable not merely to the addition of ASAT scores to the scaling criterion variable v_i, but also to the fact that those course scores scaled by "small group procedures" have in fact been fixed in consultation and with some knowledge of both ASAT and scores in other courses, and these "small group scaling procedure" scores remain fixed (cf. Remark 4.1).

It is possible to test this influence of small group scores by a jackknife study with ASAT scores rescaled so as to yield TE scores with the present standard deviation, and observe how the "imprecision" estimates in column (ii) are affected. The three rows in Table 8.3 correspond to the first three rows of Table 8.2, except that now the ASAT scores have been rescaled by a factor of 1.2 in the first two rows. The observed increases in the scale of the TE scores fell short of this 20% increase by 0.4 (i.e., 19.6% increase, because $120 - 100 \times 82.379/68.902 = 0.44 \approx 0.4$), 3.0, 4.2, 2.1, 1.5, 0.7, 1.0, 0.5, 0.6, 2.4, 1.9, 0.2, and 0.5 respectively. Comparison of the entries under (i) in Tables 8.3 and 8.2 show slight changes in these jackknife estimates of variance: approximately, 0% change

Table 8.3

Jackknife estimates of "imprecision" with rescaled ASAT scores: The effect of small group scores

Run ID #'s		SD TE Sc.	"Imprecision" variance estimates			Change from 8.2	
Tbl.8.2	This #		(i)	(ii)	(iii)	(i)	(ii)
540	561	82.379	15.937	99.821	114.034	+0.1%	+0.5%
		93.666	16.082	52.608	67.077	+0.0%	–3.0%
		79.583	18.158	61.625	76.056		
541	562	77.046	12.591	91.900	105.610	–1.9%	+8.9%
		89.402	14.676	167.589	181.926	+1.1%	–1.7%
		79.638	16.176	58.248	71.497		
542	563	54.805	13.458	81.377	94.752	–4.6%	+11.7%
		57.457	13.947	140.689	157.593	–3.4%	+16.8%
		63.281	17.286	47.036	58.619		
543	568	67.459	12.045	39.080	49.767	–6.3%	+5.2%
		76.339	11.992	43.229	55.515	–4.2%	+45.9%
		72.083	13.495	23.547	34.301		
544	564	74.180	10.270	47.556	57.767	–4.0%	+5.1%
		80.320	10.379	66.107	76.475	–2.5%	+11.3%
		74.578	13.933	33.235	44.578		
545	569	79.165	11.424	29.859	41.092	–0.9%	+1.9%
		82.740	11.672	34.518	45.720	–1.5%	+2.1%
		79.840	16.189	22.511	35.552		
546	565	77.820	10.591	39.159	49.782	+1.7%	+2.4%
		80.314	10.755	50.805	62.400	+2.5%	–5.5%
		76.614	13.500	28.685	39.771		
547	572	67.403	8.806	27.186	35.947	–4.3%	+1.3%
		75.963	8.872	35.918	44.879	–4.8%	+1.9%
		72.421	10.754	16.444	25.326		
548	566	70.453	15.996	56.277	70.860	+0.1%	+1.1%
		75.293	16.170	57.412	72.455	+1.2%	+9.3%
		77.068	17.025	33.336	46.207		
549	567	85.522	11.945	29.662	41.802	+0.0%	+6.9%
		86.948	12.166	32.918	45.689	+0.8%	+15.8%
		81.083	12.903	23.816	35.062		
550	570	65.666	14.597	32.637	47.221	–4.7%	+4.4%
		65.867	14.675	37.234	51.903	–5.9%	–2.0%
		71.549	18.075	22.205	36.867		
551	571	88.089	24.653	39.788	63.235	0.0%	+0.6%
		100.954	25.580	75.227	100.474	+0.2%	+0.5%
		90.535	26.876	25.973	48.674		
552	573	79.873	14.567	29.527	44.312	–2.8%	+2.1%
		92.183	14.692	32.739	48.166	–2.7%	–17.2%
		83.937	17.384	19.325	34.234		

in three cases, 2% increase in a fourth, and reductions of 2%, 4%, 5%, and 6% in the other four cases. *A priori*, we should expect a reduction if anything. Also, most entries in (ii) increase. Both these observations are consistent with the existing procedure placing excessive emphasis on ASAT scores ("excessive" because the noise in ASAT scores is quite the larger of these two measures).

Theory also predicts that for increasing N, $(S^2_{\mathrm{JNF}})_k$ should be $O([\#\ \text{course groups}]/N)$, and

Table 8.4

Quasi-invariant quantity (8.6) for jackknife variance estimate

Run #		ASAT Wt. = 0 course score			ASAT Wt. = 1.5 course score			ASAT
		(i)	(ii)	(iii)	(i)	(ii)	(iii)	Scal.
540	501	0.0529	0.3299	0.3771	0.0603	0.2048	0.2527	0.3248
		0.0534	0.1802	0.2272	0.0569	0.1299	0.1779	
541	503	0.0365	0.2400	0.2803	0.0460	0.1657	0.2034	0.3565
		0.0413	0.4848	0.5257	0.0508	0.2574	0.2941	
542	500	0.0496	0.2561	0.3048	0.0608	0.1654	0.2061	0.3190
		0.0508	0.4236	0.4841	0.0651	0.2391	0.2852	
543	512	0.0487	0.1408	0.1832	0.0511	0.0892	0.1299	0.2612
		0.0474	0.1122	0.1598	0.0522	0.0955	0.1410	
544	505	0.0448	0.1893	0.2337	0.0583	0.1391	0.1866	0.2485
		0.0446	0.2485	0.2938	0.0566	0.1832	0.2311	
545	507	0.0427	0.1086	0.1499	0.0600	0.0834	0.1317	0.1618
		0.0439	0.1253	0.1674	0.0566	0.0951	0.1408	
546	504	0.0461	0.1694	0.2149	0.0598	0.1270	0.1761	0.1945
		0.0464	0.2380	0.2858	0.0615	0.1553	0.2066	
547	509	0.0538	0.1568	0.2103	0.0628	0.0961	0.1480	0.2927
		0.0545	0.2059	0.2624	0.0620	0.1252	0.1789	
548	508	0.0555	0.1932	0.2436	0.0591	0.1157	0.1604	0.3070
		0.0555	0.1824	0.2339	0.0576	0.1329	0.1791	
549	506	0.0389	0.0903	0.1297	0.0420	0.0775	0.1142	0.1715
		0.0393	0.0925	0.1330	0.0430	0.0850	0.1228	
550	502	0.0663	0.1352	0.2012	0.0782	0.0961	0.1595	0.2111
		0.0674	0.1644	0.2318	0.0811	0.1074	0.1756	
551	511	0.0895	0.1435	0.2288	0.0975	0.0942	0.1766	0.3620
		0.0927	0.2716	0.3557	0.0988	0.1674	0.2513	
552	510	0.0575	0.1109	0.1697	0.0667	0.0741	0.1313	0.1986
		0.0579	0.1518	0.2126	0.0644	0.1043	0.1615	

also proportional to (1 − [corr(ASAT, TE score)]^2) × var(TE score), so a useful statistic to check this is the quantity

$$\frac{N_k(S^2_{\text{JNF}})_k}{(\#\text{ course groups})\cdot(1-\text{corr(ASAT, TE score)}]^2)\cdot\text{var(TE score)}\}}\,. \tag{8.6}$$

It is tabulated in Table 8.4 for the three resampling schemes listed at (i)–(iii) above and also for resampling using the 1985-style ASAT scaling procedure. The data shown for this last procedure should be similar to what would result had the 1986-87 procedure with its use of multiple "common scales" been used.

In Table[23] 8.4, and considering columns (i) in particular, comparison with entries in Table 8.2 suggests change from the variability pattern as there, but no obvious systematic pattern of change emerges. Nor does there appear to be any obvious relation between size of college and any of the

[23] In applying (8.6) to compile the entries for Table 8.4, the quantity used for the number of scaling groups takes no account of the "intermediate" size groups, while for convenience I used data on corr(ASAT, TE score) from the 1986 *Year 12 Study*.

entries shown. The ASAT Scaling method parameters are less variable than the others (0.162 to 0.362), but they are also the largest.

We can summarize the results of both the jackknife studies and the algebraic work concerning estimates of the scale parameters $\{\beta_j\}$ as below.

Conclusion 8.1. **The imprecision in a TE score is affected by choice of scaling procedure. Amongst procedures based on a one-factor model for the data, this imprecision is least when an Other Course Score procedure is used. The scaling parameters are model-unbiased when they are estimated by Method-of-Moments.**

Conclusion 8.2. **The major source of computational imprecision in TE score construction is associated with the use of ASAT scores. This imprecision can be considerably worsened by using the existing bivariate adjustment approach rather than the more direct estimation approach in a one-factor model for multivariate data.**

Chapter 9

The A. C. T. Tertiary Entrance Score: Background and Prescriptive Structure

The aim of this chapter is to give some background information concerning the ACT TE score which we call TEACT for the sake of distinguishing it from other actual or putative Tertiary Admission Indices. Its origins are described in Chapter 2 of *MATHEF* and are mostly not repeated here.

In terms of the discussion in Chapters 2–4 above, TEACT is a best 3.6 score aggregate. The reason for the number 3.6 stems partly from its origins in the NSW HSC score: it was a deliberate step that 3.6 should represent rather less of a student's total T-accredited curriculum load than was then used elsewhere in Australia, a situation that persists except for the aggregate score produced in Western Australia since 1985.

There are two facets to a Tertiary Admissions Index like the ACT TE score. The first covers the nature of the data available and the purpose for which the index is to be used: it is mostly this aspect that we have discussed so far.

The other concerns the prescriptive construction of the index. Such prescriptive decisions can have consequences in terms of student participation in different parts of the curriculum. For example, as a result of the 1986 changes to TE score compilation, there has been reduced participation at ACT single-sex girls' schools in those courses whose scores are scaled against ASAT-Q. As far as I am aware, this reaction is an unintended consequence of the changes introduced then: in terms of optimising TE scores it is a correct strategy. As another example, confusion over certification statements concerning *particular* as against *general* academic achievements, has altered the nature of the data now reported on the NSW Higher Schools Certificate and the construction of some Tertiary Admission Indices there since 1986 (Daley & Eyland, 1987).

Conclusion 9.1. **The use of ASAT sub-scale scores as in the 1986 ACT scaling procedure for constructing TE scores contravenes Principle P 1**.

The Source Data

The primary task perceived in this study concerns scaling procedures on the data as supplied. Consequently the school-based compilation of course scores from unit scores and operations in constructing the moderation groups from possibly more than one course, have been largely accepted without question. Put another way, we have not scrutinized the educational judgments involved in constructing the numerical summaries of the assessments, but have looked at the structural properties and processing of the resultant numerical information. This does not mean that it is not possible to elucidate some information about the properties of different constructions of course scores from studying the end-products alone, as for example in comparing the change in NSW HSC procedures in using school-based exam. mark estimates until 1985 and school-based assessments since 1986 (Daley & Eyland, 1987).

Contrary to the implied thinking of the Supervisory Committee (cf. Chapter 1 Appendix), no study of a procedure for processing numerical data can be made adequately in isolation from detailing the structural properties of the data set.

The Existing Scaling Procedure

Unless specific qualification is entailed, the term *existing scaling procedure* is used as a generic description of the first- and second-moment ASAT-equating method used in the ACT since 1977. There have been various modifications effected from time to time; those of which I am aware are listed at ASATModif. 1–8 below. [Brackets identify brief comments.]

ASATModif. 1. A truncation procedure ensuring a maximum TE score of 360 "marks" ($= 3.6 \times 100$ "marks") was used only in 1977.

ASATModif. 2. Neither ASAT nor course scores of "mature age" (MA) students are used in determining the parameters used in the scaling procedure. [The operational definition of "mature age" is questionable: in both 1986 and 1987 several students aged 21 years or more at 31 December 1986 were not classified as being of mature age.]

ASATModif. 3. Neither ASAT nor course scores of students of non-English speaking background (NESB) are used in determining the parameters used in the scaling procedure. [The definition and identification of NESB students has slowly evolved, but still does not appear to be applied uniformly for all colleges.]

ASATModif. 4. Up until *c.* 1983 a student's ASAT Total score ASAT$-T$ was determined by the number of correct items scored. Since then, it has been subject to some psychometric control in its construction [supposedly] so as to give equal weightings to the scores of the sub-tests which in terms of their face validity measure relative Quantitative and Verbal skills. [(i) Constructing the sub-scale scores entails a statistical procedure. (ii) Because of the different measurement errors in these two sub-tests, with a larger error in the Verbal sub-test where also there are usually fewer items (cf. Conclusion 5.1), the resulting ASAT-T is slightly biased towards the Quantitative sub-scale. (iii) Because of the relative numbers of items in these two areas, the gender-linked difference in ASAT-T scores has been reduced slightly by use of the post-1984 definition. (iv) It is also the case that sometimes there are items on the ASAT paper which are not classified as being in either sub-test, and under this later definition such items have been discarded from the construction of ASAT-T which thereby acquires a larger measurement error than is necessary.]

ASATModif. 5. Onwards from 1985, ASAT scores have been reported publicly on a scale with mean 150 and standard deviation 25, replacing the 1977–84 parameter values of (65, 15) . This change is purely cosmetic: an exact linear relation exists between TE scores on the pre- and post-1985 scales:

$$\begin{aligned}\text{(post-1985 TE scale score)} &= 540 + (5/3) \times (\text{pre-1985 TE scale score} - 234) \\ &= 150 + (5/3) \times (\text{pre-1985 TE scale score}). \qquad (9.1)\end{aligned}$$

ASATModif. 6. Onwards from 1986 a Writing Task was introduced and its scores AWR combined[24] with ASAT Verbal sub-scale scores to produce an *ACT Verbal score.* ASAT-T was then defined as equally weighted combinations of ASAT Quantitative sub-scale scores and ACT Verbal scores. [The major consequence of this change was to produce scores with a 10 to 20% smaller gender-linked bias between ASAT and TE scores than occurred previously, this being consistent with predictions based on evidence of the mode-of-assessment effect (cf. Daley, 1986a, b).]

ASATModif. 7. Onwards from 1986, not one but three reference scales have been used [supposedly] to provide a common scale for all courses: ASAT-T, -Q and ACT Verbal scores have been used

[24] In 1986, ACT Verbal was formed by rescaling a 70 : 30 mixture of ASAT-V and AWR to a standard deviation of 25.0. In 1987 and 1988 a 50 : 50 mixture was used.

for different courses, depending on a prescriptive decision regarding the content of the courses concerned. [The change was argued partly because it reduces the error mean square between the course and reference scale scores. It introduces a bias relative to the assumption of using a common scale implied in the construction of an aggregate score, because the courses using sub-scale scores are those attracting students with relatively higher competence in the related Quantitative or Verbal skill areas, so their mean sub-scale scores are higher than the scores of the common ASAT Total scale, and a selection bias effect ensues. The same comment would apply to the use of several reference scales as considered in Morgan & McGaw (1988). What all this bears out is M&B's remark quoted at the start of Chapter 2.]

ASATModif. 8. From time to time there have been problems associated with producing course scores in moderation groups where the numbers are "small". In the ACT the critical size for "small" has varied, generally between 5 and 10. On the other hand the size of intermediate groups has not changed at all.

Moderation Group Assumptions

In the analyses done for this study, three group sizes are of significance for the procedures adopted:

Standard Group: If the number of scores in the group is at least 20 then both first and second moments are scaled.

Small Group: If the number of scores in the group is smaller than 10 then the course scores are unchanged.

Intermediate Group: If the number of scores in the group lies in the range ≥ 10 and < 20 then the first moment is scaled statistically but the second moment is left unchanged.

The strategy for Intermediate Groups was adopted as a necessary expediency because the second moment in the data as supplied to me had already been subject to non-statistically based rescaling in accordance with the existing procedure; to attempt an analogue for an Other Course Score procedure in the absence of the original college-based data involved too much inversion at the time of programme development.

Our initial Method-of-Moment scaling operations were carried out on college data sets complete apart from omitting ASAT scores of students classified as NESB. This was a change from the existing procedure in that we used course scores of NESB students in constructing the within-college scaled scores. The change is more consistent than the existing procedure with the principle P 2 that comparative teachers' judgments are to be preserved. There was an oversight in not eliminating MA students from this scaling step, due to insufficient information about the MA indicator variable in the data as supplied. The oversight involved both the inappropriate use of ASAT scores of some MA students (cf. ASATModif. 2) and an improper construction of MA students' TE scores from rather fewer course scores than the regulation "best 3.6 scores". This oversight has subsequently been corrected so that in all analyses done for this report, the course scores used for scaling are those of the students in each college with a full TE package ($\geq$ 3.6 course units) and aged $<$ 21 on 31 December 1986. The ASAT scores of NESB and MA students are excluded. This exclusion, the definition of Group sizes, and the use of linear transformations as rescaling devices, agree with current practice.

Structure of TE Scores and Other Course Score Scaling Criteria

A student's curriculum in the ACT is made up of at least 30 units of study (an average student seeking a TE score has a load of about 6 units per term under the three-term year format). These units must be sufficiently related so as to make up at least three (or four) major and three (or one)

Put Table 9.1 on this page
[print NICOUNTS.DOC from MS-Word with PSCRPTL]

minor courses, with at least three major and one minor courses “Tertiary Accredited”. In practice, about 7% of students with a TE package have only the minimum number of T-accredited courses; all other students have at least a fourth major or two minor courses in their curriculum (see Table 9.1). Indeed, the mean number of T-accredited courses per student is around 5.0 courses for all ACT colleges (see Table 9.2).

TABLE 9.2

Mean Numbers of TE-Course Units per Student

College	Mean No. Units
cop	4.82
dck	4.86
ern	4.73
hwk	5.03
nar	5.01
phl	4.87
str	5.17
dar	4.88
edm	4.93
mar	5.19
cce	5.15
mer	5.08
stc	5.61

Some Possible Abuses

It is an assumption of almost all scaling procedures used in Australia that, across their chosen courses, students will exert approximate “parity of effort”, though this may vary between student. One is obliged to suspect that when the signal to noise ratio of scaled course scores in a college falls below about 2.0 more obvious departures from this assumption are being reflected. This reflection, like the gender-linked bias between TE and ASAT scores, lowers the correlation between TE and ASAT scores. How do various scaling procedures reflect this departure from assumed behaviour?

Under the existing system, there appears to be no penalty for students, especially the more able, to take on an extra course (even if only a minor) but not treat it so seriously. A college tends to gain an advantage from this behaviour because scaled scores in such “extra” courses are likely to be students’ lower scores, and hence lower than the scaling variables, whether ASAT or an averaged Other Course Score. As a result, the mean scaled scores of other students in the group are then inflated.

Under an Other Course Score scaling procedure, reference scores like ASAT scores, provided they are “valid for scaling” (i.e., not NESB or MA) are used just once for all students. This means that the ranking of students in the college as a whole is not inflated by more frequent use of higher ASAT scores. Amongst Other Course Score procedures, there can be inflation effects for some courses relative to the model-unbiased estimates of scale parameters $\{\beta_j\}$ associated with Method-of-Moment estimation. In the presence of smaller or larger numbers of students with “uneven” curriculum effort (cf. the previous paragraph), this optimal scaling procedure can act so as to depress slightly or more markedly all scale parameters, so that it partly compensates for the inflation effect that occurs otherwise.

Another approach to this problem of coping with an incorrect assumption involving “ethical” behaviour, is to construct the aggregate from a considerably larger part of a student’s load. Most

Australian systems do just that. In NSW for example, apart from MA students covered by special provisions, at least 10 and often 11 Units of NSW HSC study are required (roughly speaking, five majors and one minor in ACT terminology). Excluding MA students, 95% of the candidature meet this eligibility requirement, and for them, 7% offer 10 units, 52% 11 Units, 37% 12 Units, and less than 4% 13 or more Units. In WA where the number of courses required for an aggregate was reduced *c.* 1985 (and, reduced slightly more than the ACT's 3.6 course score requirement, but with an additional requirement of including both a quantitative and a humanities course score), there have been responses from students indicating course selection behaviour that is not totally in accord with the explicit and implicit "behavioural" assumptions of those specifying the new aggregate score.

Since all scaling procedures assume that every T-accredited course score should be reckoned according to its weight, so the procedures will be more or less efficacious according as students behave consistently with the "parity of effort" assumption.

CONCLUSION 9.2. **Consideration should be given to redefining a TE score as the sum of a student's best 4.5 course scores, conditional on the inclusion of at least three Major course scores, where a Minor course has a weight of 0.5 instead of 0.6, and where a Total ASAT Score made as an optimal mixture of Quantitative and Verbal sub-scales and the Writing Task, is regarded as a Minor course score and can be included in this "best 4.5 aggregate".**

Prescriptive Actions and ASAT Scores

It can hardly be stressed too strongly that the construction of any Tertiary Admissions Index like the Tertiary Entrance score is governed predominantly by the nature of the data involved. In particular, it is in general more difficult to change the nature of Tertiary Entrance scores by *prescriptive* measures concerning ASAT scores, such as were tried onwards from 1979 following on from Morgan's (1979) analyses of 1978 data demonstrating the gender-linked bias in ASAT scores relative to TE scores.

Further, *what is observed in the ACT is little different from what is observed elsewhere amongst students of similar age in other English-speaking countries* (I have not looked at literature concerning non-English speaking countries). Thus, the changes that were observed in ACT TE scores as a result of the modifications to their construction between 1985 and 1986 were largely predictable, except[25] possibly for the relative scores of students from Government and non-Government schools on

the Writing Task and the ASAT verbal sub-test (cf. Table 6.5 and Conclusion 6.4). In particular, it was both predictable and predicted that these changes would not go nearly far enough to eliminate the gender-linked bias in ASAT scores.

[25] From conversation with some individuals who have had contact with students who have completed Year 12 under the ACT system, the observed effect is not entirely unexpected. To what extent the effect would survive after reduction of the contrast factor between ASAT and course scores is conjectural at the time of writing.

Chapter 10

Some A. C. T. Course Score Analyses

The fact that any "score equating" or scaling procedure is justifiable only to the extent that data comply with positivity assumptions like those of Chapter 2 or else conform to a statistical model as in Chapter 3 implies that both the procedures and the data should be examined in relation to the model. As a compromise between this fact and the Supervisory Committee's rejection of item (i) of the Appendix to Chapter 1, this Chapter merely sketches some data analyses that firstly vindicate the preference for an Other Course Score scaling procedure over the existing ASAT scaling procedure. They also indicate, as has been observed with other data sets of secondary school students seeking a tertiary admission qualification, that a two-dimensional structure may provide an adequate description of the data set.

The analyses reported here are comprehensive in the sense of including what analyses have been done, but not in the sense of covering all courses or all colleges.

Three-factor Structure

In terms of the models developed in Chapter 3, one obvious starting point from which to consider the structure of the data set is with the *c.* 90% or more of the TE students of a college with both English and Mathematics scores. For this sub-population, confined slightly further by excluding NESB and MA students (and this exclusion applies throughout any discussion of ASAT scores), each student i has the two course scores denoted E_i and M_i, and the three scores on the ASAT Quantitative and Verbal sub-tests and the Writing Task Q_i, V_i and W_i. These five scores are independently obtained observations that are regarded as reflecting student i's school-based achievements in two "distinct" course areas and developed abilities or skills in (broadly speaking) similar areas.

In Chapter 6 we indicated analyses on these scores omitting $\{W_i\}$ supporting a three-factor representation involving a general ability *cum* achievement measure v_i, a quantitative/verbal contrast, and a mode of assessment contrast. These analyses can be amplified by the correlation matrices and 4- and 5- factor analyses listed in Table 10.1. In particular, we concluded earlier that **there exists for each student i a systematic difference between the school-based general achievement measure $v_i + \Delta_i$ and the external ASAT-determined general measure of developed ability $v_i - \Delta_i$.**

What do we learn now from the analyses summarized in Table 10.1 of the data sets $\{(E_i, M_i, V_i, Q_i)\}$ and $\{(E_i, M_i, V_i, Q_i, W_i)\}$? First, all the correlations in (a) are positive. Next, the smallest tend to be associated with either W_i or V_i, consistent with their having the largest measurement errors of the five quantities. Perhaps surprisingly, corr(E, M) is rather smaller than corr(V, Q). Since both pairs contrast quantitative and verbal skills, interpret this as showing that the contrast Δ_i of Chapter 6 originates more in ASAT than school-based assessment. Also, corr(W, E) is rather larger than corr(V, E): this underlines that the skills on which ASAT-V draws are more distinct than those used by E and W.

From table (b) giving the latent roots, we conclude that at least two but no more than three factors are discernible. The first factor, whether in (c) or (d), corresponds to the general achievement *cum* ability factor discussed in Chapters 2–4. Note that the English scores load least strongly of the four components. The second factor in (c) is predominantly a contrast between Q and E,

Table 10.1

Correlations, Latent Roots, and Factor Loadings of
Q, V, W, E and M Scores in 1986 within the 13 ACT Colleges

(a) Correlations

College	Q/V	Q/W	V/W	Q/E	V/E	W/E	Q/M	V/M	W/M	E/M
1	0.577	0.215	0.336	0.292	0.385	0.450	0.689	0.472	0.231	0.481
2	0.667	0.279	0.346	0.458	0.503	0.476	0.784	0.577	0.270	0.561
3	0.483	0.243	0.367	0.320	0.414	0.601	0.646	0.326	0.311	0.529
4	0.550	0.355	0.416	0.280	0.300	0.521	0.722	0.378	0.277	0.478
5	0.656	0.368	0.432	0.279	0.419	0.527	0.700	0.511	0.418	0.473
6	0.689	0.329	0.420	0.447	0.510	0.581	0.673	0.441	0.343	0.491
7	0.634	0.111	0.385	0.252	0.453	0.599	0.644	0.388	0.217	0.400
8	0.634	0.292	0.356	0.316	0.458	0.573	0.645	0.472	0.265	0.503
9	0.679	0.426	0.356	0.576	0.601	0.471	0.639	0.397	0.239	0.604
a	0.652	0.489	0.368	0.561	0.576	0.674	0.646	0.436	0.433	0.643
b	0.659	0.210	0.338	0.538	0.606	0.576	0.745	0.607	0.177	0.510
c	0.712	0.460	0.483	0.380	0.523	0.628	0.678	0.586	0.423	0.631
d	0.677	0.313	0.290	0.530	0.555	0.468	0.706	0.484	0.290	0.608

(b) Latent roots of 4- and 5-component correlation matrices

1	2.467	0.740	0.543	0.250		2.680	0.987	0.615	0.471	0.247
2	2.787	0.588	0.441	0.185		3.013	0.901	0.488	0.415	0.183
3	2.368	0.686	0.694*	0.252		2.704	0.954	0.690	0.409	0.242
4	2.382	0.787	0.623	0.208		2.722	0.932	0.721	0.434	0.191
5	2.542	0.755	0.491	0.212		2.925	0.886	0.516	0.461	0.212
6	2.632	0.596	0.558	0.213		2.980	0.853	0.558	0.396	0.213
7	2.402	0.781	0.605	0.212		2.649	1.165	0.606	0.372	0.208
8	2.523	0.706	0.526	0.244		2.818	0.966	0.562	0.432	0.222
9	2.752	0.429	0.605	0.213		3.031	0.783	0.581	0.414	0.191
a	2.759	0.431	0.577*	0.234		3.202	0.702	0.546	0.369	0.180
b	2.837	0.552	0.362	0.249		3.040	1.023	0.377	0.311	0.249
c	2.763	0.655	0.391	0.191		3.207	0.770	0.510	0.351	0.161
d	2.783	0.480	0.529*	0.208		3.012	0.837	0.522	0.425	0.202

* For the sake of demonstrating common structure of the second and third factors, and to maintain correspondence with the table of factor loadings, the eigenvalues here are not in descending order.

with any alignment of M and V being more with Q than E. The third factor is a contrast between V and M, as is hardly surprising in view of the second, with no discernible patterns of alignment of the other two quantities. Coupled with our analyses of Chapter 6, what these imply is that a quantitative/verbal contrast makes two appearances, and is here confounded with a mode of assessment contrast. (Note that in 4 of the 13 colleges, we have identified the second factor with the factor loadings of the third largest latent root.)

Turning to table (d), the second factor is now unashamedly a contrast between Q and M on the one hand (with maybe a little help from V), and E and W on the other. The third factor is a contrast between V (with weak support from Q) and M generally supported by E, while W flips between the external multiple-choice measures and the school-based measures. We interpret these two factors as contrasts of quantitative/verbal domains and multiple choice/school-based assessments respectively.

Table 10.1 (cont.)

(c) Factor loadings in analyses of correlation matrices of Q, V, E, M

College	First factor				Second factor				Third factor			
	Q	V	E	M	Q	V	E	M	Q	V	E	M
1	0.838	0.777	0.652	0.858	0.399	0.159	–0.745	0.033	0.157	–0.595	–0.025	0.405
2	0.885	0.824	0.736	0.885	0.332	0.118	–0.671	0.117	0.126	–0.545	0.029	0.357
3	0.811	0.704	0.724	0.831	0.387	0.377	–0.604	–0.171	0.321	–0.582	–0.276	0.420*
4	0.860	0.714	0.626	0.861	0.319	0.369	–0.735	–0.091	0.250	–0.584	–0.235	0.405
5	0.853	0.820	0.639	0.855	0.404	0.128	–0.758	0.041	0.075	–0.530	–0.025	0.452
6	0.878	0.818	0.738	0.805	0.340	0.020	–0.659	0.213	–0.073	–0.527	0.112	0.513
7	0.844	0.807	0.639	0.793	0.428	–0.058	–0.744	0.203	–0.086	–0.543	0.138	0.532
8	0.833	0.812	0.691	0.832	0.439	0.149	–0.701	–0.002	0.072	–0.522	–0.067	0.494
9	0.878	0.807	0.839	0.792	0.383	0.016	–0.520	0.110	–0.064	–0.535	0.055	0.559*
a	0.864	0.798	0.837	0.821	0.402	–0.099	–0.482	0.164	–0.119	–0.549	0.176	0.479*
b	0.880	0.854	0.777	0.854	0.283	–0.134	–0.570	0.360	0.038	–0.494	0.268	0.211
c	0.842	0.853	0.748	0.875	0.444	0.229	–0.627	–0.114	0.132	–0.430	–0.139	0.411
d	0.880	0.812	0.801	0.842	0.330	–0.174	–0.501	0.300	–0.155	–0.521	0.304	0.376*

*See footnote to table of eigenvalues (= latent roots) at (b).

(d) Factor loadings in analyses of correlation matrices of Q, V, W, E, M

First factor

	Q	V	W	E	M
1	0.792	0.771	0.554	0.696	0.817
2	0.853	0.814	0.554	0.766	0.854
3	0.736	0.694	0.674	0.787	0.779
4	0.812	0.715	0.674	0.681	0.795
5	0.805	0.799	0.700	0.684	0.825
6	0.828	0.802	0.670	0.781	0.770
7	0.750	0.803	0.601	0.733	0.738
8	0.785	0.790	0.632	0.751	0.783
9	0.869	0.794	0.606	0.845	0.751
a	0.841	0.756	0.735	0.868	0.794
b	0.841	0.845	0.530	0.825	0.811
c	0.813	0.831	0.735	0.785	0.836
d	0.856	0.792	0.556	0.821	0.819

	Second factor					Third factor				
	Q	V	W	E	M	Q	V	W	E	M
1	0.453	0.116	–0.700	–0.436	0.297	0.110	0.433	0.303	–0.473	–0.317
2	0.351	0.129	–0.760	–0.310	0.297	0.141	0.332	0.262	–0.510	–0.171
3	0.560	0.056	–0.607	–0.391	0.340	0.030	0.686	–0.011	–0.207	–0.420
4	0.462	0.129	–0.562	–0.522	0.337	0.017	0.579	0.249	–0.382	–0.422
6	0.433	0.149	–0.633	–0.379	0.315	0.081	0.527	0.003	–0.128	–0.508
7	0.567	0.095	–0.685	–0.485	0.360	0.108	0.535	0.024	–0.190	–0.521
8	0.461	0.198	–0.646	–0.457	0.297	0.142	0.433	0.212	–0.282	–0.480
9	0.131	0.004	–0.746	–0.010	0.457	0.127	0.566	–0.230	–0.125	–0.419
a	0.299	0.459	–0.583	–0.243	0.051	–0.064	0.408	0.221	0.041	–0.570
b	0.353	0.092	–0.789	–0.335	0.393	–0.125	0.500	–0.149	0.005	–0.300
c	0.460	0.266	–0.490	–0.471	0.161	0.166	0.200	0.404	–0.295	–0.439
d	0.284	0.217	–0.794	–0.181	0.213	0.041	0.531	0.067	–0.138	–0.464

Two-factor Structure of Course Scores

Having indicated the nature of much, if not most, of the information available in course and ASAT scores, it would now be proper to consider the nature of information available in course scores alone. In this report we simply outline some *ad hoc* analyses done in order to understand the relevance or otherwise of the representation at equation (3.14), namely that student i's scaled course score Y_{ij} in subject j is expressible as

$$Y_{ij} = v_{i1} + \gamma_j v_{i2} + e'_{ij}\,. \tag{10.1}$$

As an aside that we do not follow up here, there are at least two simple-minded independent estimators of the factor v_{i2} that we could consider: subject to suitable scaling of the scores, for most students we could use either or both of $Q_i - V_i$ and $M_i - E_i$.

Our major aim here is to consider a variety of linear predictors using various combinations of the five scores in $\{(E_i,\, M_i,\, V_i,\, Q_i,\, W_i)\}$. This is equivalent to studying linear regressions in terms of these variables for each set of course scores. Since the set of course scores is constant for each of these regressions, comparisons of different regressors within sets can be effected by noting the proportion of unexplained variance (i.e., in the jargon of multivariate analysis, the quantity $100 \times (1 - R^2)\%$), and this is what is listed in Table 10.2. The following generalizations are made on the basis of the various analyses summarized in this table:

(a) Use of the estimator $\frac{1}{2}(V_i - Q_i)$ of v_{i2} in conjunction with $T_i \equiv \frac{1}{2}(V_i + Q_i)$ tends to add little explanatory power except for Mathematics course scores. This is not to say that no quantitative/verbal contrast factor v_{i2} is discernible, because using the course score estimator $\frac{1}{2}(E_i - M_i)$ of v_{i2} in conjunction with the analogue $\frac{1}{2}(E_i + M_i)$ of T_i leads to somewhat increased explanatory power in several cases. The problem with the ASAT sub-scale scores is probably tied in with both the size of their measurement and model-fit errors and the presence of the factor Δ_i.

(b) The set $\{(E_i,\, M_i)\}$, or even just $\{\frac{1}{2}(E_i + M_i)\}$, is generally markedly superior to the set $\{(V_i,\, Q_i,\, W_i)\}$, entailing much more than the reduction in unexplained variability that has been seen so far in coming from any estimate of the contrast factor v_{i2}. The better performance of $\{(E_i,\, M_i)\}$ is again understandable in terms of the presence of the factor $\{\Delta_i\}$ in the latter set but not the former in relation to sets of other course scores, and of the relative sizes of measurement errors.

(c) The Other Course Score estimators $\{v_i\}$ at (3.9) mostly have lower unexplained variance (equivalently, higher correlation: see *Year 12 Study*) than even a best-fitting linear combination of $\{(E_i,\, M_i)\}$. This reflects the decreased measurement and model-fit error in v_i relative to the pair $(E_i,\, M_i)$.

(d) To the extent that it is relevant, the contrast factor v_{i2} is not as well estimated by either or both of the pairs $(V_i,\, Q_i)$, $(E_i,\, M_i)$ as the general measure $v_i + \gamma'_i |v_{i2}|$. Accordingly, the estimation of e.g. General Quantitative and Verbal skills by estimators of $v_i \pm v_{i2}$ almost certainly entails appreciable loss of precision over estimation of a general achievement measure. Rather *careful* predictive validity studies are needed to check on the relative worth of any aggregates as predictors. Educationally, the use of more than one aggregate conflicts with P 1.

(e) In spite of having larger measurement errors and cruder lattice span imprecision than the ASAT scores, the set of scores $\{W_i\}$ provides increased explanatory power, though not to the same extent as using course scores. In general the scores W_i replace and add to whatever contribution comes from V_i, but this is hardly surprising!

From the tables of correlation coefficients in Tables 12 of the annual *Year 12 Study* publications of the ACT Schools Accrediting Agency, we can make allowance for the inclusion of some course scores in TE scores and thereby deduce that

TABLE 10.2

Various Linear Regressors Compared as Predictors of Course Scores via Unexplained Variance

College and Course	T	$\{V, Q\}$	$\{V,Q,W\}$	$(E+M)$	$\{E, M\}$	
A English	51.6%	49.1%	40.2%		$\{$ACV, $M\}$: 40.0%	$\{V, Q, M\}$: 39.4% $\{V,Q,W,M\}$: 33.1%
A Mathematics	62.0%	57.8%	56.5%	Q: 58.3%		$\{Q, E\}$: 46.8% $\{V,Q,W,E\}$: 45.1%
A Physical Science	34.8%	34.8%	31.3%	13.3%	12.1%	
A Biological Sci.	82.5%	82.1%	61.3%	38.3%	38.3%	
A Humanities (i)	51.4%	50.7%	44.3%	45.8%	37.5%	
A Humanities (ii)	67.8%	64.0%	51.4%	37.2%	28.9%	
B Physics	74.0%	73.0%	72.0%	43.7%	42.7%	
B Chemistry	56.5%	52.3%	49.0%	42.1%	30.1%	
B Biological Sci.	65.0%	64.9%	62.5%	43.0%	43.0%	
B Computing	84.6%	84.1%	81.1%	50.2%	48.2%	
B Drama	68.1%	67.6%	52.9%	38.7%	30.9%	
C Mathematics	43.8%	37.6%	37.5%	Q: 38.4%		$\{Q, E\}$: 33.0%
D Mathematics	67.8%	58.8%	58.8%	Q: 59.1%		$\{Q, E\}$: 50.9%
						$\{E{+}M, V,Q,W\}$
E Drama	97.1%	*c.*96%	91.5%	80.1%	79.0%	72.3%
E Biological Sci.	65.3%	64.7%	61.2%	56.4%	55.1%	47.1%
E Social Sci.	68.2%	66.0%	58.6%	33.1%	33.1%	28.5%
E History	74.3%	74.2%	74.1%	34.9%	31.3%	33.8%
E Art	77.3%	75.1%	55.4%	48.1%	36.7%	38.7%

Code for scores: V, Q = ASAT sub-scale, W = Writing Task. $T = \frac{1}{2}(V + Q)$. ACV = ACT Verbal $= 0.7V + 0.3W$. E = English, M = Mathematics, $E + M$ = analogue of T.

$$\text{corr(TE score, course score)} > \text{corr(ASAT score, course score)}.$$

We have also checked more extensive analyses of 1984 data along the lines of Table 5.2 of *MATHEF*. These three sets of overlapping analyses are mutually consistent.

CONCLUSION 10.1. **Other Course Score scaling procedures have considerably smaller error mean squares than ASAT scaling procedures, and are therefore in general superior procedures.**

Since parameter unbiasedness is a desirable property, this in turn points to using a procedure having at the very least the model-unbiasedness properties of the Method-of-Moment estimators of the scaling parameters (cf. Daley, 1987a, 1988). The empirical work that has been sketched in this section also supports the implications of this algebraic work and the conclusions of Chapters 2 and 3: no matter what higher dimensional sub-space may be common to course scores as distinct from any "uniqueness" factors, approximate unbiasedness of estimation of the dominant general achievement measure comes from using the Method-of-Moment estimation procedure.

§5.9 of *MATHEF* quotes p.100 of M&B that "the correlations of the Australian Scholastic Aptitude Test with some course scores are clearly too low to justify current attempts to bring these scores to a common scale using the Aptitude Test as a reference test" (i.e., rejection of A 5). Table 10.2 bears that out (and here we ignore the fallacy of a common scale for multivariate data set: see the discussion below (4.29)). Tables 10.2 and 10.1 and Chapter 6 identify the ASAT scores as a source of a multidimensional problem that is irrelevant to a discussion of course scores. At the very least they make highly questionable M&B's contention that "the obvious multidimensionality of the course [scores] renders invalid any attempt to summarize student performance in a single score", because the most obvious source of deviation from unidimensionality that emerges from Table 10.2 is the ASAT scores which explain less of the variability than the Other Course Scores.

ASAT scores no longer have any place as such in the ACT Tertiary Entrance Statement. Their role has always been ancillary to the school-based assessments under the ACT guidelines, supposedly providing merely system-wide estimates of developed academic ability for use in constructing group estimators (recall A 5). It follows that any deviation between the measures of developed academic ability (or whatever ASAT scores measure) and relative academic achievement as determined by school-based assessments, should be identified so soon as it is seen as a cause of problems in producing any Tertiary Admissions Index. Such problems occurred directly in the 1977 query about the gender-linked bias and indirectly in the relatively low correlation even with aggregates of course scores. Chapter 6 and Tables 10.1–2 establish the existence of the contrast factor Δ_i. There is an urgent need to reduce its influence.

Chapter 11

An Optimally Constructed A C T Tertiary Entrance Score

This Chapter details the numerical steps needed to produce a TE score in the ACT starting from a set of course scores $\mathcal{X}$ and using the Optimal Other Course Score Scaling Procedure (colloquially, "Method-of-Moment" scaling). Focus attention for the moment on one college and write

$$\begin{aligned}\mathcal{X} &= \{X_{ij} : i = 1, \dots, N;\ j \text{ in set } \mathcal{S}_i \text{ of courses where } i \text{ has scores}\},\\ \mathcal{W} &= \{w_{ij} : \text{weight of course } j \text{ for student } i\},\end{aligned}$$

where each non-zero w_{ij} takes one of the values 0.6, 1.0, 1.6, 2.0, according as the course score is a Minor, Major, Major-Minor, or Double Major. (Strictly, for ACT use at present, our use of the symbol j is ambiguous, sometimes denoting a course and sometimes a *moderation group*, because it is possible for a student to have more than one course score in the same moderation group, as for example English and Media Studies when the latter school-based scores are put in the English moderation group.) Depending on the number N_j of scores in group j, its set of school-based scores is treated by small group scaling procedure (I have followed 1986 ACT practice and taken $N_j \leq 9$ for this), intermediate group procedure ($10 \leq N_j \leq 19$), or standard group procedure ($N_j \geq 20$).

For a *small group*, scores are determined by a consultation process that includes reference to ASAT scores and other course scores; no automated scaling procedure is involved. This means that for j in a small group, scaled scores Y_{ij} are determined before and independently of any statistical scaling procedure.

For the other groups, an automated procedure is employed, with slightly different details for intermediate and standard groups so far as estimates of the scale factors β_j are concerned. For the moment regard all groups as either standard or small. The aim of the exercise is to find estimators $(a_j,\, b_j)$ of the scaling parameters $(\alpha_j,\, \beta_j)$ and construct the *scaled scores*

$$Y_{ij} = a_j + b_j X_{ij}\,. \tag{11.1}$$

An iterative procedure is employed. Use a superscript as in $a_j^{(n)}$ to denote values at the n^{th} step. For convenience, and it is desirable but not essential, start with scores in $\mathcal{X}$ that have a common mean and standard deviation, (150, 25) say, for each moderation group. Write $\mathcal{S}_i$ for the set of groups where i has scores X_{ij} (equivalently, those j for which $w_{ij} > 0$), and $\mathcal{C}_j$ for the set of individuals with scores in group j (when individuals have more than one score in group j, they are counted for as many scores as they have in group j). Start from

$$(a_j^{(0)},\, b_j^{(0)}) = (0,\, 1). \tag{11.2}$$

For $n = 0, 1, \dots$ and $i = 1, \dots, N$ construct

$$v_i^{(n)} = \frac{\sum_{j \text{ in } \mathcal{S}_i} w_{ij}(a_j^{(n)} + b_j^{(n)} X_{ij})}{\sum_{j \text{ in } \mathcal{S}_i} w_{ij}}\,. \tag{11.3}$$

For j in moderation groups at the given college, excluding small groups but including group 0 (see below (11.6)), define

$$\mathrm{ave}_j(v_i^{(n)}) \equiv \frac{\sum_{i \text{ in } \mathcal{C}_j} w_{ij} v_i^{(n)}}{\sum_{i \text{ in } \mathcal{C}_j} w_{ij}}\,, \tag{11.4}$$

$$\mathrm{var}_j(v_i^{(n)}) \equiv \frac{\sum_{i \text{ in } \mathcal{C}_j} w_{ij} (v_i^{(n)} - \mathrm{ave}_j(v_i^{(n)}))^2}{\sum_{i \text{ in } \mathcal{C}_j} w_{ij}}\,, \tag{11.5}$$

$$\mathrm{cov}_j(X_{ij},\, v_i^{(n)}) \equiv \frac{\sum_{i \text{ in } \mathcal{C}_j} w_{ij} X_{ij} (v_i^{(n)} - \mathrm{ave}_j(v_i^{(n)}))}{\sum_{i \text{ in } \mathcal{C}_j} w_{ij}}\,, \tag{11.6}$$

where the reference scale scores and associated weights are designated as group 0 (these weights will typically be either 0 for NESB and MA students, and some positive constant otherwise: use $w_{i0} = 0.000001$, say, for reference scale weight of 0). Define for the moment

$$b_0^{(n+1)} = \frac{\mathrm{var}_0(v_i^{(n)})}{\mathrm{cov}_0(X_{i0},\, v_i^{(n)})} \tag{11.7}$$

and then for $j \neq 0$ and not a small or intermediate group, set

$$b_j^{(n+1)} = \frac{1}{b_0^{(n+1)}} \cdot \frac{\mathrm{var}_j(v_i^{(n)})}{\mathrm{cov}_j(X_{ij},\, v_i^{(n)})} \tag{11.7}$$

(if j is a small or intermediate group, $b_j^{(n)} = 1.0$ for all n). Now reset

$$b_0^{(n+1)} = 1. \tag{11.9}$$

Compute $\mathrm{ave}_j(X_{ij})$ and $\mathrm{var}_j(X_{ij})$ analogously to (11.4) and (11.5) (in practice, this is done once and for all at the outset). Momentarily set

$$a_0^{(n+1)} = \mathrm{ave}_0(v_i^{(n)}) - \mathrm{ave}_0(X_{i0}), \tag{11.10}$$

in order to determine for $j \neq 0$ and not in a small group,[26]

$$a_j^{(n+1)} = \mathrm{ave}_j(v_i^{(n)}) - b_j^{(n+1)}\, \mathrm{ave}_j(X_{ij}) + a_0^{(n+1)}\,. \tag{11.11}$$

Now reset

$$a_0^{(n+1)} = 0. \tag{11.12}$$

The system of equations from (11.3) onwards can now be iterated until convergence is attained. This will typically take up to 10 or 12 iterations for convergence of the intermediate values $(a_0^{(n+1)},\, b_0^{(n+1)})$ to their putative values (0, 1). The eventual proximity of these values to (0, 1) is a measure of the success of a scaling procedure (recall from the discussion preceding Conclusion 2.3 that for a balanced data set $\mathcal{X}$, convergence to (0, 1) would be attained: in practice the data set is not balanced, and we have instead a statistical model as in Chapter 3).

[26] The original has – instead of + for the term $a_0^{(n+1)}$.

It remains to fix suitable weights $\{w_{i0}\}$ for the reference scale scores. In this report I have always used $w_{i0} = 0$ for Mature Age or NESB students; understand in the sequel that these exceptions have been made. In some analyses in this report I have used $w_{i0} = 1.5$ because this value yields Tertiary Entrance scores TEMM with about the same variance as the existing TE scores TEACT. After doing several of these analyses, I changed to using $w_{i0} = 0.8$ as being closer to what substitution in equation (4.29) yields for ACT data, coupled with rescaling of the ASAT scores by a factor $110/103(= 110/(95 + 10 \times 0.8))$ because this preserves to within 1 or 2% the existing TE score variance in the scores TEMM; this is desirable on account of the Small Group scaling procedures (see discussion preceding (4.21)).

A third parameter should be used to reduce the effect of outlier scores (see the end of Chapter 4). Finally, the ASAT reference scale scores should be subjected to a gender-linked calibration to eliminate the sex-bias effect. This done, the process should be iterated, but unlike the two or three iterations required of the calibration procedure sketched in *MATHEF*, I should expect one iteration here to suffice because the effects on TE scores of gender-linked changes to ASAT scores are now one step more remote from those ASAT scores than under the existing scaling procedure.

CONCLUSION 11.1. **On the basis of what is presently known, there are four steps required to construct a TE score as fairly as possible via a set of linear transformations of school-based scores, consistent with the Basic Assumptions and Principles for Constructing TE scores:**

(1) Determination of scaling parameters via Method-of-Moment estimation using Other Course Scores as the basis of the scaling criterion variables, in conjunction with (2)–(4) below.

(2) Removal of the gender-linked bias in ASAT scores.

(3) Reduction of the effects of outlier scores (this may overlap with (1)).

(4) Fixing suitable weights for ASAT- (or whatever-) based reference scale scores in relation to non-statistically determined course scores, as for example with small groups.

I have not detailed here the analogous formulae to be followed for intermediate size moderation groups, as I have not had the raw school-based data from which to work. Briefly, what is involved is to determine the scale parameters by *OptOCSP* and use this as a component of a mixture as ASAT scores are presently used for such groups.

Chapter 12

Some Comparisons of Three Scaling Procedures via TE Scores and other Parameters

The aim of this Chapter is to illustrate some consequences of using different scaling procedures. We start by giving three sets of "Tertiary Entrance Scores" denoted TEACT, TASAT and TEMM. They are constructed from course scores obtained from three different scaling procedures. We give examples of the different rankings that come from using these different TE scores, and illustrate ways of summarizing such information. The scaling procedures differ in the scaling parameters they produce for the various moderation groups. We list these sets of parameters coming from the various colleges classified according to the "course areas" of the scores in the moderation groups. The only systematic differences that emerge between the procedures and that are linked to the different areas, are explicable in terms of the properties of the procedures that are known *a priori*: we find no other evidence of systematic effects being introduced via the Optimal Other Course Score Scaling Procedure (*OptOCSP*).

Some Examples of Different TE Scores

TEACT, TASAT and TEMM result respectively from applying the 1986 scaling procedure using ASAT Total and sub-scale scores, from the 1985 scaling procedure using just ASAT-T scores, and from *OptOCSP* discussed in Chapters 3, 4 and 11 using Method-of-Moment estimation with ASAT weighting factor WTAS = 0.8 and a rescaling factor on ASAT-T scores of 110/103 to maintain approximately the same standard deviation of TE scores as for TEACT or TASAT.

Table 12.1 lists the three scores for all students with just the minimum number of course scores needed to construct a TE score, i.e., 3.6 units. To lessen the identifiability of particular students' scores, the three scores for each student have been adjusted by an integer between –9 and +9, allocated randomly to each set of scores as listed.

Table 12.2 is similar to Table 12.1 except that now the entries are grouped by college and are of students with a common number of units. Their scores cover most of the range of TE scores, and have been shifted as before to lessen identifiability.

These tables show that changes in TE scores of up to 20 or 30 points can come simply from changing the scaling procedure (e.g., in Table 12.1, students with TEACT scores in the range 500 to 504 have TASAT scores from 485.8 to 505.4, and TEMM scores from 476.8 to 532.8). These larger changes are not the norm, but they are far from uncommon. What are the effects of these changes on student rankings?

Table 12.1

Sets of 1986 (TE)ACT Scores, (T)ASAT-Scaled Scores, and Method-of-Moment (TE)MM Scores: All Students with 3.6 Units

ACT	ASAT	MM	ACT	ASAT	MM	ACT	ASAT	MM	ACT	ASAT	MM
353	353.4	368.5	355	353.2	362.1	357	350.4	367.0	367	364.9	383.7
371	369.5	391.1	379	373.4	356.2	379	374.7	359.3	381	365.6	385.1
386	372.8	338.7	390	390.7	396.7	393	388.9	404.0	395	393.1	406.3
396	391.3	392.0	397	393.4	395.8	402	392.7	404.4	402	414.4	414.7
406	404.1	407.0	407	400.1	409.2	411	396.2	410.6	413	410.5	415.3
414	403.0	430.4	416	417.5	443.4	416	419.0	421.6	418	411.0	417.6
420	419.9	427.5	420	420.0	419.7	421	420.0	424.2	421	420.7	427.8
423	413.5	394.1	423	418.2	432.8	423	420.1	427.8	424	423.3	428.6
426	419.4	427.8	427	424.8	432.5	430	419.4	437.5	434	434.0	437.2
435	424.3	435.9	435	432.6	432.1	435	432.9	433.9	435	434.8	439.5
436	427.7	430.6	436	429.4	432.3	437	435.0	445.2	440	437.6	443.7
441	434.0	432.2	443	436.8	438.3	444	425.2	428.3	444	452.5	460.4
445	435.6	434.7	447	435.8	436.2	447	442.8	454.7	448	435.0	458.9
450	436.9	446.5	450	443.8	449.6	450	445.7	448.1	450	446.8	448.8
451	444.5	446.4	451	449.5	458.0	451	450.3	445.2	452	451.5	457.1
453	447.7	456.7	455	453.6	461.7	455	453.9	446.3	455	455.1	458.3
457	455.1	458.3	458	457.6	459.1	459	455.4	466.7	459	456.1	461.4
461	452.0	453.2	461	457.1	459.5	462	455.2	462.6	462	461.7	464.4
463	449.9	455.4	463	453.2	453.9	463	457.3	468.9	468	465.0	466.3
469	460.4	463.3	469	465.4	466.9	469	466.2	479.2	469	467.3	471.1
470	465.3	472.3	471	463.0	459.3	473	464.8	468.0	473	471.7	477.4
473	472.2	470.0	473	474.0	477.4	473	474.2	485.3	474	470.0	462.4
474	470.8	471.7	475	476.1	479.3	477	470.9	470.4	477	478.4	477.5
478	475.0	484.7	481	479.5	496.7	481	481.6	489.8	482	480.5	484.9
483	480.6	484.0	484	474.6	464.8	484	477.4	476.3	484	481.5	484.4
485	481.9	487.6	486	480.1	486.8	486	484.6	488.5	487	478.0	477.4
487	484.4	483.3	488	483.6	481.1	490	484.4	493.0	490	488.3	488.1
490	488.4	490.0	491	479.9	475.3	491	489.1	492.0	491	489.9	495.4
494	482.3	510.2	494	491.4	496.4	495	493.7	497.2	496	493.3	486.8
496	497.0	492.0	497	494.4	499.0	497	495.2	497.6	499	496.0	496.2
499	496.6	497.6	500	488.7	489.6	500	500.6	508.1	501	485.8	476.8
502	494.8	492.1	502	499.7	495.6	502	502.7	500.9	503	490.4	486.8
504	505.4	532.8	505	502.1	498.0	507	503.3	505.9	508	498.0	497.2
511	511.0	517.3	512	502.2	504.4	517	507.0	504.2	518	510.9	510.4
519	515.1	505.8	522	514.0	517.9	522	518.5	526.5	522	522.2	530.1
527	514.2	525.9	528	527.2	529.5	529	522.2	516.6	531	520.6	523.2
531	526.7	532.0	531	527.0	532.9	533	534.0	530.5	535	526.2	521.6
538	541.0	546.1	539	536.1	533.7	541	540.7	547.8	546	520.7	518.6
546	544.9	547.5	550	545.7	550.5	551	547.1	548.2	554	546.0	546.5
564	562.7	554.3	566	559.1	571.9	567	565.5	566.4	576	572.5	569.4
578	571.0	572.2	581	577.8	573.4	581	580.5	606.6	583	578.0	575.7
600	595.4	601.4	603	598.3	598.5						

TABLE 12.2

Sets of TE Scores:

Students with the Same Number of Units Grouped by College

ACT	ASAT	MM	ACT	ASAT	MM	ACT	ASAT	MM	ACT	ASAT	MM
College A											
438	431.8	436.8	445	440.5	424.4	479	454.5	454.3	479	467.4	483.0
485	473.6	483.2	500	492.7	484.8	503	497.0	504.7	513	503.6	512.4
520	493.7	498.7	532	526.0	521.7	532	526.1	531.4	534	524.8	540.5
536	531.4	528.1	542	535.7	539.6	545	539.7	555.8	548	535.8	538.6
554	541.0	553.0	556	545.9	548.1	556	547.8	543.1	559	555.0	560.2
560	556.0	550.0	564	561.4	556.6	575	570.4	571.2	577	568.1	574.9
590	582.7	579.5	593	591.6	596.7	595	589.3	584.7	596	590.3	603.7
601	596.8	610.9	609	604.1	621.4	613	610.2	615.9	619	616.1	621.2
620	619.0	616.7	643	639.1	641.0	644	641.2	646.4	647	643.4	645.9
649	641.2	656.4	651	646.6	653.6	654	647.5	654.2	663	661.0	668.1
680	675.3	681.4	698	694.9	699.2						
College B											
415	405.9	430.9	417	422.5	427.3	485	476.9	480.9	490	489.6	493.6
493	492.3	487.4	493	494.6	497.6	505	504.3	505.9	509	505.8	506.6
511	504.9	503.1	520	507.8	503.0	520	519.6	521.2	527	519.9	519.6
532	521.0	520.8	532	526.6	528.3	540	537.4	533.4	544	533.5	531.0
550	533.5	529.9	550	533.8	527.2	552	549.4	549.6	555	547.4	549.3
558	542.7	536.6	562	558.9	556.8	569	565.9	561.2	570	560.3	558.7
571	569.3	566.7	579	569.4	564.1	579	574.6	574.2	591	580.5	575.8
592	587.1	584.1	598	595.7	597.2	600	591.4	585.5	602	596.0	585.9
606	596.6	591.4	613	608.9	606.0	617	611.9	609.7	617	613.3	611.0
619	612.9	613.7	630	619.6	614.5	630	626.2	626.6	634	625.7	620.7
635	631.8	629.8	644	637.1	631.8	668	661.6	654.9	694	687.3	677.2
697	694.0	685.6	697	695.7	680.5	714	710.0	700.6			
College C											
334	330.6	335.4	367	365.1	377.1	432	426.8	433.5	437	430.8	444.4
453	448.7	457.9	457	452.6	459.8	465	459.2	464.7	471	449.9	431.3
483	476.1	485.7	485	482.9	491.6	487	484.0	495.5	491	487.8	495.8
491	488.0	495.9	493	490.9	489.6	505	490.9	488.5	507	500.6	505.6
510	501.5	504.0	521	514.0	513.2	525	519.5	524.8	535	533.3	536.1
538	536.5	540.7	539	533.1	530.5	542	538.0	537.4	542	539.6	540.5
551	546.3	554.3	551	546.4	557.0	554	550.2	554.3	555	551.2	556.8
556	554.3	557.1	557	553.5	557.8	557	553.5	561.9	558	555.8	553.0
560	556.7	561.5	560	557.9	565.4	561	556.4	565.3	563	557.7	556.8
564	560.9	562.0	565	558.9	545.9	565	561.7	565.9	570	562.3	552.7
570	566.7	571.2	571	566.1	566.9	573	562.1	551.9	575	568.0	558.4
575	570.9	581.9	578	575.7	576.1	580	574.2	569.5	580	578.3	582.8
581	570.6	562.7	582	580.6	573.1	583	581.2	584.1	586	582.9	574.2
586	584.7	582.6	591	582.4	576.9	594	589.4	582.8	608	603.5	601.7
613	609.0	594.7	624	619.4	606.8	627	625.1	616.0	632	629.9	633.7
641	639.3	634.4	653	653.8	650.2	654	653.6	642.1	662	663.3	650.9
668	666.1	671.1	674	676.9	666.0	675	673.7	690.4	676	679.4	669.3
681	682.7	669.3	698	696.3	712.0	699	698.5	696.4	715	720.5	715.0
725	723.2	728.1									

Table 12.2 (cont.)

College D

481	478.1	482.9	497	489.1	488.6	508	501.6	499.6	525	522.9	529.6
526	521.9	517.8	526	523.5	521.5	529	514.1	516.2	531	529.3	542.6
538	530.6	538.0	543	540.8	538.0	552	533.4	537.1	561	549.6	553.8
563	550.0	557.1	565	561.9	561.3	568	564.0	557.4	570	569.5	582.0
574	571.4	566.4	576	573.7	581.2	583	567.2	572.5	583	573.0	575.8
585	570.6	572.6	599	586.9	589.2	599	592.8	603.0	600	592.4	594.8
603	595.0	587.9	603	603.5	602.8	611	610.8	602.2	612	601.1	603.8
614	608.5	613.5	616	615.4	606.7	621	620.2	621.1	622	621.8	637.2
626	618.8	616.7	628	617.6	617.8	638	626.9	626.0	648	641.2	638.4
648	648.0	652.6	651	643.6	643.3	657	654.0	643.8	659	655.2	647.3
659	655.5	650.0	672	665.0	659.9	678	670.6	669.6	680	673.5	669.7
680	675.1	671.2	681	675.1	671.3	686	678.5	673.4	699	691.1	684.6

College E

485	484.0	480.6	505	492.0	481.6	519	493.0	477.3	533	530.9	534.8
540	542.0	541.5	556	555.5	551.4	562	564.1	558.6	581	579.5	582.9
606	595.2	584.6	617	603.5	595.4						

College F

463	461.5	473.1	465	460.8	467.6	472	470.0	470.2	489	472.9	464.9
521	519.3	516.5	524	516.2	514.7	541	539.3	543.3	559	559.5	552.8
562	555.2	553.2	565	557.0	546.0	575	571.4	568.1	581	563.2	552.5
586	577.2	566.8	586	578.7	579.7	586	580.9	574.4	589	592.7	586.8
598	582.6	575.1	599	584.5	577.7	600	593.8	590.2	610	614.2	604.1
618	605.0	592.6	625	615.8	610.0	641	626.1	618.9	657	647.5	640.8
680	683.4	669.6									

College G

424	420.7	413.7	425	422.1	406.8	426	423.4	416.6	430	427.4	438.6
434	431.0	439.5	451	452.1	438.4	455	453.0	446.8	466	463.3	451.2
466	463.3	463.8	468	465.6	470.9	472	471.0	474.7	473	470.5	458.8
480	474.7	482.1	481	478.2	484.7	483	472.6	484.2	483	485.5	490.8
492	487.4	493.0	492	488.0	480.5	496	483.1	494.8	498	489.4	478.2
506	501.1	491.4	507	503.3	522.8	517	515.2	522.5	519	520.8	521.6
519	522.1	526.8	524	526.9	516.6	527	530.0	530.4	531	530.5	531.8
534	519.2	529.4	535	530.7	528.3	535	534.3	537.9	538	534.6	531.4
538	540.6	539.1	540	539.9	535.2	543	542.8	546.9	547	550.2	535.5
550	548.0	533.7	554	554.1	554.3	558	555.2	542.5	562	558.2	560.8
565	563.1	556.5	565	567.2	572.3	573	559.8	568.7	573	574.1	573.8
574	575.1	567.8	578	578.2	580.7	580	577.7	583.5	582	568.4	576.9
587	585.4	582.1	590	590.5	579.4	597	597.7	611.4	597	600.5	593.4
600	599.0	598.0	603	599.7	606.7	604	602.9	596.2	618	618.2	617.7
622	627.4	619.6	626	625.2	642.2	632	633.4	624.5	634	638.8	628.1
637	635.0	619.3	642	640.1	631.8	643	639.2	633.6	657	660.7	645.5
660	649.2	649.6	668	665.9	649.7	676	675.0	659.9	689	687.3	677.7

College H

465	430.7	426.7	492	479.4	483.5	544	532.1	533.5	549	526.9	517.4
550	547.8	545.1	576	554.7	543.8	579	562.5	553.2	590	593.9	603.3
592	588.8	596.1	606	608.3	606.0	612	605.9	604.5	618	604.5	595.5
622	618.6	603.2	625	613.3	608.3						

Table 12.2 (cont.)

College I

425	419.1	414.0	465	462.8	463.0	487	471.0	463.4	488	480.0	480.5
495	492.6	502.9	505	503.1	501.3	507	501.5	494.6	508	484.8	478.0
511	508.5	512.4	516	512.7	516.4	518	505.6	495.6	519	494.6	482.2
520	513.8	507.3	520	515.1	515.0	539	537.1	546.6	559	548.2	542.7
564	561.7	559.0	567	563.4	568.7	570	560.7	559.3	573	569.8	564.5
575	568.6	559.8	585	582.1	580.3	592	589.6	600.3	594	582.9	575.1
599	595.2	592.9	655	653.7	642.6						

College J

511	508.2	503.7	513	517.8	516.3	516	509.6	516.7	517	525.4	519.7
522	535.2	525.3	560	556.3	561.4	564	575.9	577.6	607	598.4	599.8

College K

410	400.2	400.9	425	416.3	402.0	426	428.6	421.5	472	464.2	463.4
472	486.9	476.9	479	481.4	481.0	488	493.5	490.2	510	513.4	506.3
570	579.2	572.1	585	581.8	577.0	606	624.4	619.3	608	607.0	610.3

College L

461	456.5	457.3	519	507.2	508.6	540	539.4	537.8	564	547.7	546.9

College M

448	447.1	451.7	458	455.4	461.0	471	470.1	485.7	487	484.6	488.4
499	497.8	494.0	510	508.7	508.9	521	513.7	507.7	527	525.8	524.7
534	533.0	530.8	543	538.7	546.2	547	545.4	546.1	550	549.2	541.0
550	550.5	546.8	552	549.2	555.3	556	555.5	550.1	561	557.4	554.4
568	567.5	567.7	570	570.2	569.8	579	578.8	571.1	588	587.8	580.6
590	591.2	591.8	592	592.5	587.6	593	594.0	590.5	601	600.9	599.8
601	601.0	603.9	603	604.3	591.7	605	604.5	601.6	610	611.3	603.8
616	615.8	619.5	619	618.2	619.4	623	623.5	616.1	630	630.8	629.4
641	640.9	629.9	663	665.9	665.5	669	669.2	664.6	669	669.3	667.7
669	671.3	672.0									

Rankings from Different TE Scores

We show the effects of the changes on students with scores at the other end of the scale from Table 12.1. Table 12.3 is a 3-in-1 table: the three sets of three columns correspond to the top 99 students ranked according to the criteria TEACT, TASAT, and TEMM respectively. Each set of columns gives the rankings according to TEACT (column 1), TASAT (column 2) and TEMM respectively.

The first 99 students ranked by any criterion constitute about the top 4% of students. Depending on what is taken as the base, so the set of students included in the top 99 changes. For example, if we take TASAT as the base set of scores, corresponding more or less to the 1985 TE scores, then the use of sub-scales as in 1986 to yield TEACT scores would mean exchanging 2 students in the top 9, or else 6 students in the top 99. If instead *OptOCSP* is used so as to yield TEMM scores, then 1 student in the top 9, or 10 students in the top 99, would be changed. Finally, the use of TEMM rather than TEACT (or *vice versa*), would lead to 2 students in the top 9, or 14 in the top 99, being affected by the change.

The arguments developed in this report point to the use of a single reference scale (hence, to TASAT or TEMM), and to Other Course Score scaling (hence, to TEMM), so the implication of the sizes of these changes is that TEACT is an even shoddier scale than TASAT relative to TEMM.

Table 12.3

Students' TE Score Rankings from Different Scaling Procedures

Ranking by: TEACT			TASAT			TEMM		
TEACT	TASAT	TEMM	TEACT	TASAT	TEMM	TEACT	TASAT	TEMM
	Rankings			Rankings			Rankings	
1	1	2	1	1	2	2	2	1
2	2	1	2	2	1	1	1	2
3	3	3	3	3	3	3	3	3
4	5	4	5	4	6	4	5	4
5	4	6	4	5	4	22	6	5
6	22	30	22	6	5	5	4	6
7	24	33	8	7	8	10	12	7
8	7	8	9	8	9	8	7	8
9	8	9	19	9	23	9	8	9
10	12	7	16	10	10	16	10	10
11	13	14	17	11	11	17	11	11
12	16	13	10	12	7	18	14	12
13	17	19	11	13	14	12	16	13
14	21	21	18	14	12	11	13	14
15	25	34	23	15	31	25	23	15
16	10	10	12	16	13	24	19	16
17	11	11	13	17	19	55	46	17
18	14	12	27	18	20	44	30	18
19	9	23	24	19	16	13	17	19
20	20	25	20	20	25	27	18	20
21	27	26	14	21	21	14	21	21
22	6	5	6	22	30	32	29	22
23	15	31	25	23	15	19	9	23
24	19	16	7	24	33	29	34	24
25	23	15	15	25	34	20	20	25
26	43	47	34	26	36	21	27	26
27	18	20	21	27	26	30	49	27
28	28	29	28	28	29	35	41	28
29	34	24	32	29	22	28	28	29
30	49	27	44	30	18	6	22	30
31	33	40	39	31	38	23	15	31
32	29	22	52	32	48	33	47	32
33	47	32	31	33	40	7	24	33
34	26	36	29	34	24	15	25	34
35	41	28	57	35	56	43	38	35
36	54	52	63	36	54	34	26	36
37	51	55	58	37	58	49	45	37
38	52	41	43	38	35	39	31	38
39	31	38	41	39	51	117	115	39
40	50	43	45	40	46	31	33	40
41	39	51	35	41	28	38	52	41
42	62	63	53	42	49	48	44	42
43	38	35	26	43	47	40	50	43
44	30	18	48	44	42	94	69	44
45	40	46	49	45	37	76	78	45
46	55	82	55	46	17	45	40	46
47	58	57	33	47	32	26	43	47
48	44	42	66	48	66	52	32	48

Table 12.3 (cont.)

TEACT	TASAT	TEMM	TEACT	TASAT	TEMM	TEACT	TASAT	TEMM
49	45	37	30	49	27	53	42	49
50	68	65	40	50	43	51	53	50
51	53	50	37	51	55	41	39	51
52	32	48	38	52	41	36	54	52
53	42	49	51	53	50	110	83	53
54	64	64	36	54	52	63	36	54
55	46	17	46	55	82	37	51	55
56	74	73	74	56	78	57	35	56
57	35	56	64	57	60	47	58	57
58	37	58	47	58	57	58	37	58
59	63	72	81	59	80	82	72	59
60	66	69	65	60	70	64	57	60
61	67	62	68	61	79	102	108	61
62	80	77	42	62	63	61	67	62
63	36	54	59	63	72	42	62	63
64	57	60	54	64	64	54	64	64
65	60	70	69	65	83	50	68	65
66	48	66	60	66	69	66	48	66
67	88	95	61	67	62	90	82	67
68	61	79	50	68	65	75	77	68
69	65	83	94	69	44	60	66	69
70	84	81	89	70	101	65	60	70
71	95	89	95	71	91	118	81	71
72	98	92	82	72	59	59	63	72
73	113	143	96	73	102	56	74	73
74	56	78	56	74	73	142	133	74
75	77	68	83	75	94	111	85	75
76	78	45	97	76	106	92	96	76
77	87	103	75	77	68	62	80	77
78	89	98	76	78	45	74	56	78
79	101	110	98	79	111	68	61	79
80	99	105	62	80	77	81	59	80
81	59	80	118	81	71	70	84	81
82	72	59	90	82	67	46	55	82
83	75	94	110	83	53	69	65	83
84	110	104	70	84	81	85	86	84
85	86	84	111	85	75	146	158	85
86	90	86	85	86	84	86	90	86
87	92	126	77	87	103	101	100	87
88	93	149	67	88	95	185	202	88
89	70	101	78	89	98	71	95	89
90	82	67	86	90	86	168	169	90
91	104	114	119	91	124	95	71	91
92	96	76	87	92	126	72	98	92
93	124	117	88	93	149	126	146	93
94	69	44	104	94	122	83	75	94
95	71	91	71	95	89	67	88	95
96	73	102	92	96	76	128	97	96
97	76	106	128	97	96	107	143	97
98	79	111	72	98	92	78	89	98
99	107	147	80	99	105	143	157	99

TABLE 12.4

Amended Rankings from Changing TE Scores

Top Proportion:	50%	20%	10%	4%
(a) Amended Student Rankings				
Increase				
5 points	47.8%	18.5%	9.1%	3.5%
15 points	43.4%	15.6%	7.4%	2.8%
25 points	39.0%	13.2%	5.9%	2.1%
(b) Approximate Proportionate Change in Group				
3 points	1.1%	1.9%	2.3%	2.9%
5 points	1.8%	3.1%	3.9%	4.8%
7 points	2.5%	4.3%	5.5%	6.7%
9 points	3.2%	5.6%	7.0%	8.6%
11 points	3.9%	6.8%	8.6%	10.5%
(c) Proportionate Change of Group from Bias				
Bias from "True"				
5 points	4.4%	7.6%	8.8%	11.4%
15 points	13.2%	21.2%	26.2%	31.0%
25 points	21.9%	34.2%	40.5%	46.8%

Further, recall that none of these procedures involves removal of the gender-linked bias: we should expect the ultimate changes to be even higher still (cf. Conclusion 11.1).

CONCLUSION 12.1. **The use of a particular scaling procedure can have considerable influence on the set of students meeting a TE score based selection criterion, particularly in the more selective groups. In comparison with the OptOCSP, the 1986–88 procedure is most noticeably discrepant, even before the removal of the gender-linked bias.**

Another way of illustrating the effects of the changes is to use distributional properties and differences of pairs of TE scores. It is sufficient for these purposes to approximate the distribution by a normal distribution (the use of an arbitrary distribution is sketched in Daley, 1987a). Suppose that the scores TEMM and TEACT have the same means and standard deviations, with the latter equal to 90. What is the effect on a student's ranking of a difference in the two TE scores of 5 or 15 or 25 points? Table 12.4[27] shows the amended rankings at different points in the range; note that if the standard deviation is smaller than 90 then the effects as shown would be greater still.

Part (a) of the table is easily interpreted: a student ranked at 50.0% of all students and whose score increases 15 points under a different scaling procedure, would move up the rankings to 43.4%. This part of the table can also be read inversely. The shift in rankings upwards of 11% such as occurred between 1976 and 1977 (see Table 7.1), is equivalent to a TE score change in 1986 TE scale points of about 25 points or more.

Part (b) of the table uses the standard deviation of TE score changes as an average measure of the change, and lists the expected proportion of any "top group" that would be affected by the TE score changes. Table 12.7 lists these standard deviations for the colleges for the three pairs of TE scores being considered here. Since the size of the ACT TE score population is about

[27] Equation (12.1) and the entries in Table 12.4(b) are corrected from the original which used 2.15 in (12.1) instead of $0.861 = 2.15/\sqrt{2\pi}$.

TABLE 12.5

Sex Bias Measures from Different Scaling Procedures

College	Discrepancy between ASAT-T and			Standard Errors		
	TEACT	TASAT	TEMM	TEACT	TASAT	TEMM
COP	22.57	27.76	31.60	9.94	10.05	10.11
DAR	4.56	3.84	4.85	10.23	10.26	10.30
DCK	13.04	15.50	17.49	8.32	8.33	8.37
ERN	43.69	51.64	58.51	13.40	13.32	13.74
HWK	16.08	17.07	23.63	9.21	9.36	9.41
NAR	4.06	4.36	7.47	8.90	9.07	9.38
PHL	15.78	18.43	19.96	9.72	9.79	9.77
STR	34.04	37.03	43.72	11.77	11.87	11.97
Weighted Mean		18.98	22.57			

2,400, the group of “top 99” students corresponds roughly to the top 4%, so the data 6/99, 10/99 and 14/99 deduced from Table 12.3 correspond to entries in Table 12.4(b). To determine the SD entries, take the median values of the standard deviations from Table 12.7, namely 5.9, 6.4 and 9.0, giving expected proportions of about 14%, 15% and 21% respectively. While the data point to the possibility of overestimation (and this is likely, in view of the nature of the changes in the different colleges), they are of the correct order of magnitude.

The entries for part (b) are computed from an approximation in Daley (1987a). Express the standard deviation between for example TEMM and TEACT scores as a fraction σ of their standard deviation 90 (say). Now ask what proportion of students in the top $100p$% we can expect to be changed as a result of using one selection criterion rather than the other. Making normal distribution assumptions (these are not necessary: see Daley, 1987a), what corresponds for example to “14/99” ≈ 0.14, is approximately

$$\frac{\sigma}{\sqrt{2\pi}} \cdot \frac{\varphi\big(\Phi^{-1}(1-p)\big)}{p}\bigg|_{p=.04} = \frac{\sigma}{\sqrt{2\pi}} \cdot \frac{\varphi\big(\Phi^{-1}(1-0.04)\big)}{0.04} = 0.861\sigma \qquad (12.1)$$

where Φ and φ denote the normal distribution function and its density.

Part (c) describes how the composition of a group of eligible students is affected by similar shifts for a (small) sub-group of students, and in particular therefore, by shifts that are biases like the gender-linked bias that holds in the ACT and Queensland. For example, if there is a gender-linked bias that depresses TE scores at a single-sex girls’ school by 15 points, that school can expect to have about 31% fewer students meeting the cutoff level to a tertiary course where only students with scores in the top 4% ranking are eligible.

Note that what is listed as “Bias from true” in Table 12.4(c) corresponds to half the bias measure as listed in Table 12.5, or *c.* 1.8 times the discrepancies listed in Table 7.4. (In the standardized notation used in equation (7.1), “Bias from True”/90 = b.) The sex bias measures in Table 12.5 between ASAT-T and each of TASAT and TEMM are conceptual analogues of the quantities in Daley (1985); they become numerically analogous, and on the same scale as there, on division by $(5/3) \times 3.6 = 6$. The measures between TEACT and ASAT-T are not conceptually analogous: the ASAT-T measure must be replaced by a mixture of ASAT-T, -Q and ACT Verbal (ACV) scores so as to reflect the bias in each of these three scales relative to the course scores (cf. Table 6.3). Since ASAT-T would be replaced more often by -Q than by ACV, this would generally tend to increase the measures shown. In other words, the TEACT/ASAT discrepancy underestimates the bias measure required for calibration.

Table 12.6

Means and Standard Deviations of TE Scores from Different Scaling Procedures

College	Means				Standard Deviations			
	TEACT	TASAT	TEMM	ASAT	TEACT	TASAT	TEMM	ASAT
	(a) Students with TE Package (Mature Age excluded)							
COP	540.37	535.59	533.94	511.54	77.23	75.82	71.93	92.68
DAR	558.92	558.40	558.16	541.75	80.36	82.47	80.11	90.76
DCK	551.41	546.10	545.44	523.36	77.85	76.77	73.29	90.81
ERN	553.51	547.36	544.12	521.74	67.31	67.03	63.64	85.53
HWK	567.06	563.60	563.84	537.45	86.58	89.08	84.83	100.94
NAR	571.32	566.97	568.07	522.98	86.72	91.70	90.47	110.37
PHL	576.49	572.22	572.48	553.73	73.94	73.70	70.22	85.96
STR	554.24	549.02	547.68	526.11	76.57	78.96	77.02	91.43
CCE	563.07	562.36	560.95	540.79	73.79	76.26	74.92	86.62
MER	537.40	539.27	536.55	508.88	78.92	82.95	81.17	86.06
STC	555.39	554.08	553.50	535.75	80.53	81.40	82.91	86.09
EDM	561.67	556.62	555.85	533.34	82.16	83.99	79.51	96.44
MAR	566.65	561.79	561.46	543.17	79.89	79.46	77.45	85.65
	(b) Students with TE Package (Mature Age and NESB excluded)							
COP	541.12	536.62	535.22	520.72	77.50	76.05	72.13	85.72
DAR	559.60	559.10	558.85	542.54	80.15	82.24	79.86	90.55
DCK	558.12	552.78	552.00	536.81	77.01	75.93	72.37	84.05
ERN	553.88	548.19	544.77	532.93	69.07	68.58	65.19	83.06
HWK	568.10	564.73	565.17	545.81	87.16	89.74	85.66	97.95
NAR	569.28	564.90	565.95	542.01	84.22	88.23	88.60	99.93
PHL	577.47	573.27	573.50	557.07	74.22	73.90	70.36	84.43
STR	560.33	555.28	553.94	534.82	72.02	74.31	72.32	87.70
CCE	565.71	564.86	563.48	546.13	73.50	75.97	74.49	86.65
MER	531.27	531.81	529.75	515.98	80.19	83.19	81.84	87.38
STC	561.49	559.39	559.12	540.66	78.81	80.06	81.90	86.00
EDM	563.87	559.64	558.53	542.88	81.75	83.10	78.65	92.55
MAR	567.79	563.06	562.70	545.45	80.40	79.90	77.87	85.41

Moments of Different TE Score Distributions

Table 12.6 lists the means and standard deviations of the three TE scores we have considered for each of the ACT colleges, and Table 12.7 lists the means and standard deviations of the differences of the three pairs of scores (the first three columns of Table 12.6(a) yield the Mean entries in Table 12.7). We see immediately from Table 12.7, or else the second and third columns of Table 12.6, that, excluding DAR College, the use of sub-scale scores in the 1986 scaling procedure relative to the 1985 procedure resulted in a gender-linked shift, with mean TE scores remaining roughly constant at the single-sex girls' schools while all other schools saw a general upwards shift of about 4 points, as predicted in Daley (1986b).

These effects can be traced to selection effects associated with using sub-scale scores and to the more frequent use of ASAT-Q scores for males (taking Mathematics, Physics and Chemistry) than of ASAT-V scores for females (mostly, just English). To see this, take Mathematics for example. Relative to ASAT-T scores, the ASAT-Q scores of Mathematics students will tend to be higher (lower) for the students who are stronger (weaker) in Mathematics. The stronger students tend to take Mathematics at a level with a higher unit count than other students, so the weighted mean

TABLE 12.7

Means and Standard Deviations of Differences of "TE" Scores Students with the Same Number of Units Grouped by College

	Means			Standard Deviations		
	(1) – (3)	(1) – (2)	(2) – (3)	(1) – (3)	(1) – (2)	(2) – (3)
COP	6.425	4.782	1.643	10.991	5.904	6.331
DAR	0.762	0.516	0.246	5.211	2.668	5.251
DCK	5.976	5.313	0.662	9.009	4.688	6.383
ERN	9.397	6.153	3.244	13.327	6.574	9.741
HWK	3.224	3.462	–0.238	8.851	4.413	8.544
NAR	3.258	4.355	–1.097	10.017	6.980	10.603
PHL	4.014	4.274	–0.260	7.364	4.354	5.866
STR	6.560	5.220	1.340	10.078	5.121	7.407
CCE	2.114	0.703	1.411	9.092	5.260	8.051
MER	0.857	–1.870	2.727	7.194	9.871	5.189
STC	1.885	1.308	0.577	8.783	7.860	5.741
EDM	5.823	5.058	0.765	12.882	7.175	8.992
MAR	5.182	4.851	0.330	7.144	6.841	3.313

Code: (1) = TEACT, (2) = TASAT, (3) = TEMM.

ASAT-Q score of all Mathematics students is higher than their mean ASAT-T score. The same is true in Physics and Chemistry which were also scaled against ASAT-Q in 1986.

According to folklore, the students opting for a curriculum with a larger component of Mathematics, Physics and Chemistry, are regarded as having TE scores possibly higher than they ought. Yet, the effects of the changes[28] instituted in 1986 were to make these scores on average higher still. Moreover, the changes depressed the mean TE scores at single-sex girls' schools where it was also agreed that the TE scores if anything were lower than should be the case.

This selection effect serves to emphasize one of the flaws of using more than one reference scale as absolute rather than relative scaling devices. It is an immediate consequence that a group of students will collectively maximize their TE scores when each student chooses courses which are scaled against the reference scale with his/her highest reference scale score. Thus, on average, under the 1986 scaling scheme, more males should tend to choose Mathematics, Physics and Chemistry, and females should tend to choose English and Drama. Adoption of the latter strategy, in the sense of girls at single-sex schools tending to move out of Mathematics, Physics and Chemistry courses, has been observed by staff at those schools.

I can only infer that it was *presumed* that using a series of reference scales, each with the same mean and standard deviation, should provide a satisfactory basis for scaling. Yet, such is not the case, nor is it consistent with how scores are treated in New South Wales for example. There, public examination marks in different courses do double service, as assessments and as reference scale scores for scaling school-based assessments in those courses. What corresponds to our data set $\mathcal{X}$ is the set of averages in each course of the exam. mark and the scaled school-based assessment. It is this analogue that is scaled via an Other Course Score procedure, and finally an aggregate is constructed from these scaled scores. In contrast the 1986 ACT procedure has multiple applications of a bivariate adjustment procedure in a multivariate data setting: it is not an appropriate approach to the construction of a TE score.

[28] Both these predicted effects of the 1986 changes were advised to the ACT Schools Accrediting Agency, which for 1986-88 opted not to amend the scaling procedure adopted in haste in response to non-technically based recommendations in *MATHEF*.

Table 12.8

Correlations of TE and ASAT Scores:
Students with TE package (excluding Mature Age NESB)

College	(1,2)	(1,3)	(2,3)	(1,4)	(2,4)	(3,4)	(1,5)	(3,5)
COP	0.99734	0.99218	0.99781	0.6638	0.6465	0.6261	0.6548	0.6448
DAR	0.99979	0.99786	0.99831	0.6932	0.6962	0.6892	0.6339	0.6338
DCK	0.99828	0.99472	0.99747	0.6756	0.6695	0.6536	0.6046	0.5919
ERN	0.99571	0.98079	0.99054	0.6406	0.6264	0.5762	0.5991	0.5889
HWK	0.99918	0.99519	0.99645	0.6718	0.6684	0.6496	0.5536	0.5524
NAR	0.99866	0.99490	0.99370	0.7355	0.7320	0.7118	0.5350	0.5379
PHL	0.99832	0.99617	0.99783	0.6172	0.6092	0.5958	0.5986	0.5895
STR	0.99827	0.99094	0.99556	0.6450	0.6408	0.6190	0.5494	0.5373
CCE	0.99796	0.99270	0.99452	0.7113	0.7170	0.6908	0.3689	0.3809
MER	0.99504	0.99701	0.99847	0.7611	0.7721	0.7607	0.4689	0.4992
STC	0.99572	0.99494	0.99773	0.7544	0.7705	0.7586	0.6053	0.6007
EDM	0.99724	0.98895	0.99528	0.7079	0.7015	0.6752	0.5948	0.5810
MAR	0.99640	0.99646	0.99943	0.7310	0.7285	0.7265	0.6190	0.6130

Code: (1) = TEACT, (2) = TASAT, (3) = TEMM, (4) = ASAT, (5) = NTI.

Conclusion 12.2. **Using more than one reference scale implies that optimal subject choices can increase a student's TE score via statistical properties of a scaling procedure. Such choices may be contrary to "educationally desirable" curriculum construction**.

A standard method of describing agreement between two sets of scores that have "much in common" is via correlation coefficients. In the present context these are listed for the three TE scores in Table 12.8. With one exception they exceed 0.99. For interest we also show the correlations of the TE scores with ASAT-T, and for TEACT and TEMM, with the number of course units (cf. Table 10.1).

TE scores measure first and foremost the quantities we have denoted $\{v_i\}$ which for the ACT have a standard deviation s_v in the region of 90. The differences in TE scores from different scaling procedures, if measured as standard deviations of (say) TEACT − TEMM, yield quantities s_Δ about 5 to 10 as shown in Table 12.7. The correlation coefficient is approximately equal to

$$\frac{s_v^2}{s_v^2 + s_\Delta^2} \approx \frac{90^2}{90^2 + 10^2} \approx 0.988.$$

Significantly, the one correlation coefficient in Table 12.8 smaller than 0.99 corresponds to a much smaller term s_v. Differences between students' TE scores within a college are directly summarized by the standard deviations listed in the right-hand part of Table 12.7.

As also shown by the entries in Table 12.4, the standard deviations of TE score changes are a more useful way of summarizing the effects on TE scores of different scaling procedures.

Conclusion 12.3. **Correlation coefficients do not usefully summarize differences in TE scores resulting from different scaling procedures.**

Regression Study Showing TEACT/TEMM Difference

Table 12.9 shows the proportions of sub-scale and Writing Task scores in the optimal regression predictors of TEACT and of TEMM computed with zero ASAT weight in the scaling criterion

TABLE 12.9

Analyses of ASAT Scores as Predictors of TE Scores

		TEACT				TEMM			Diff'ce
Run #	corr	Regression coefficients			corr	Regression coefficients			corr.
		Q :	V :	W		Q :	V :	W	coeff.s
250	0.691	56.5 :	7.1 :	36.4	0.618	41.6 :	7.8 :	50.7	0.073
251	0.811	34.2 :	21.9 :	43.9	0.807	36.2 :	20.6 :	43.1	0.004†
252	0.713	54.9 :	10.4 :	34.8	0.693	50.3 :	11.2 :	38.5	0.020
253	0.716	52.8 :	25.3 :	21.9	0.695	58.8 :	19.8 :	21.5	0.021
254	0.697	59.4 :	21.3 :	19.3	0.690	58.0 :	19.6 :	22.5	0.007†
255	0.644	48.9 :	27.5 :	23.6	0.618	39.3 :	31.1 :	29.6	0.026
256	0.758	59.0 :	9.3 :	31.7	0.732	55.8 :	9.7 :	34.5	0.026
257	0.720	34.5 :	34.2 :	31.4	0.694	24.9 :	38.9 :	36.2	0.026
258	0.686	58.2 :	13.0 :	28.9	0.634	51.6 :	13.4 :	35.0	0.052
259	0.777	42.6 :	27.1 :	30.3	0.754	44.1 :	23.7 :	32.1	0.023
260	0.634	51.6 :	9.4 :	39.0	0.615	48.7 :	10.7 :	40.7	0.019
261	0.690	46.1 :	17.3 :	36.6	0.664	41.0 :	18.4 :	40.6	0.026
262	0.738	59.6 :	16.8 :	23.6	0.729	56.8 :	18.5 :	24.8	0.009†

† See text

variables (the data set differs slightly from Table 4.1). It is immediately evident that correlations of TEACT are higher than of TEMM, and we quickly see why: the group scaling parameters are much influenced by the error components of ASAT scores in all but the largest of the moderation groups. Since ASAT scores are convex mixtures of the predictor variables, their optimal mixture then has a larger correlation with an aggregate score, or selected aggregate score. The increase in the correlation coefficient is significant for all cases except the three marked (†). The increase is associated with greater weight to Writing Task scores and usually a slightly larger fall in the weight in ASAT-Q scores, though the two are not invariably associated. The balance is necessarily reflected in ASAT-V scores where there are more rises than falls (4 falls, 9 rises). This change is a result of a statistical artefact: the increase in predictability comes purely from using a scaling procedure biased towards producing the larger correlation.

Moderation Group Parameters

A scaling procedure is effected by transforming the scores in groups. It follows then that scaling procedures can be compared at the most basic level by comparing the transformations within these *moderation* groups (as they are called in the ACT).

Tables 12.10 and 12.11 list two pairs of parameters $(a_j,\ b_j)$ needed to transform the scaled course scores X_{ij} used for constructing TEACT into the scores Y_{ij} used for TEMM (or, TASAT) via $Y_{ij} = a_j + b_j X_{ij}$. Table 12.10 comes from all mixed-sex colleges, Table 12.11 from all single-sex colleges. Each row corresponds to one moderation group. Rows are grouped according to the course area of the group (or of the lowest course area number when the group has courses from more than one area). All moderation groups, whether standard, intermediate or small, are listed.

Start with the right-hand pairs, which effectively compare the 1986 and 1985 scaling procedures. When a pair equals or is almost equal to (0, 1), it is mostly the case that ASAT-T has been used as the reference scale score under both procedures. This occurs for most entries for areas numbered 22 and up. Small deviations from (0.000, 1.000) represent the effects of numerical rounding errors (in general, I have worked and reported more decimal places to reduce rounding errors in any subsequent computations), and should be ignored. Occasionally the entries represent

First of 4 pp. for Table 12.10

Second of 4 pp. for Table 12.10

Third of 4 pp. for Table 12.10

Fourth of 4 pp. for Table 12.10

First of 2 pp. for Table 12.11

Second of 2 pp. for Table 12.11

more than just rounding error: I used an age-based definition of Mature Age student, and this differs slightly from the Agency's definition (in 1986, several students aged more than 21 years on 31 December were included in the group whose ASAT scores were used in the scaling procedure).

Most entries for areas numbered up to 20 correspond to use of ASAT sub-scale scores under the 1986 procedure. These entries reflect both slightly higher correlations of scores with the subscale rather than ASAT-T (shown here by the parameter b_j for ASAT-T scaling being larger than 1.000), and either the selection effect noted a couple of pages earlier (for ASAT-T scaling the parameter a_j tends to be negative at mixed-sex colleges) or the gender-linked difference in the Quantitative/Verbal areas (for ASAT-T scaling the parameter a_j tends to be negative for English at male single-sex colleges and for Mathematics, Computing, Physics and Chemistry at female single-sex colleges, and positive otherwise).

Now refer to the left-hand pairs. They compare the 1986 and the Optimal Other Course Score (*OptOCSP*) scaling procedures. When a pair equals or is almost equal to (0, 1), it now usually indicates a Small Group, or if only $b_j \approx 1.000$, an Intermediate Group.

Look first at the scale parameters b_j. These tend to be approximately similar to the ASAT-T parameters in groups numbered up to 20 where both procedures use a single reference scale relative to the 1986 procedure, somewhat smaller in more traditional academic course areas 22 to 62, and if anything larger in not so traditional course areas numbered upwards from 70.

Look next at the location parameters a_j. Since in general we have tried to preserve the standard deviation of TE scores between different procedures, overall reductions in scale parameters b_j must be accompanied by larger variations in the location parameters, and this is exactly what is observed. In groups with course areas numbered up to 20, the selection and gender-linked effects of the 1986 procedure relative to a single reference scale such as with ASAT-T scaling are no longer obvious. Within an area (i.e., across colleges), the variation in both scale and location parameters reflects sampling variability, indicative of the precision or otherwise with which the parameters can be determined. The only exception I may see to this statement concerns area 16. Here it is arguably a possibility that the more developed skills of problem solving *cum* analytical reasoning that advantage a student taking an ASAT paper relative to a general measure of academic achievement as is used in scaling under *OptOCSP*, exacerbate the selection bias of the scores under the 1986 procedure.

CONCLUSION 12.4. **No systematic bias effects are observable in the moderation group parameters computed via the Optimal Other Course Score Scaling Procedure**.

Chapter 13

Miscellanea

This Chapter starts by indicating how simulation ("Monte Carlo") studies of data sets may further help in evaluating the validity or otherwise of different scaling procedures. The data set used to illustrate some of this work reinforces support for *OptCSP* against *ASATSP*. The work is mainly exploratory (cf. Appendix to Chapter 1).

We conclude by noting that the approach used in this report for producing a single aggregate applies equally well to the construction of each of a set of more than one aggregate should they be so prescribed.

Scaling Procedures and Algorithms

For the purposes of implementation, a scaling procedure is an algorithm. Any computational algorithm must be specifiable in algebraic terms. In this language it should be possible to see the precise assumptions on which the algorithm is based. The collection of all these assumptions constitutes a model. And the model provides the natural reference framework for discussion of the assumptions, the algorithm, and any applications of the model to data.

The existing ACT scaling procedure is certainly specified as an algorithm, in the sense of giving a recipe for extracting one set of numbers from of another set. But as a model, and in terms of its relation to applications (e.g., in Queensland as well as the ACT), the literature is markedly deficient. The expositions in McGaw (1977) and analogous papers emerging from the ACT around the same period (e.g. Keeves, McBryde & Bennett, 1977) provide descriptions of the mechanics of the algorithm, but the scant evaluations of the assumptions they give are confined to first-order effects: the fact that second-order properties of the model are explicitly involved in the algorithm is ignored, as are the inconsistencies between ASAT and course scores. There are notes in McGaw (1987), overlapping in part with discussion in the ACT that preceded M&B, reflecting a growing awareness of this oversight, though McGaw's account still gives no hint of the fact that the problem of producing an algorithm that takes explicit account of second-order properties had been addressed at least as early as 1985 (Daley & Seneta, 1986).

For present purposes it is enough to consider the 1977–85 algorithm. Our conclusions mostly apply equally to the 1986–88 version in spite of its being inconsistent with the model describing the data.

Why Simulation Studies?

The transition from the representations of Chapter 2 to the model of Chapter 3 enables us to make the further transition from the balanced data set of Chapter 2 to the unbalanced data sets that occur in practice. We can study the latter theoretically via approximations and asymptotics as in Daley (1987a, 1988), and empirically via simulation studies, provided we can emulate the structure of such data sets.

One of the difficulties of studying data sets like those of an ACT college concerns the patterns of student course choice behaviour. These patterns are certainly associated with the course score behaviour as shown by general achievement measures v_i or by ASAT scores A_i (cf. *ASAT and TE Scores*, 1988). The shuffling routines used below destroy those associations, and so provide some evidence that any interaction effect between the scaling procedure and this dependence between general measures and course choice behaviour is at most weak. (This does not ensure that no biases come as a result! — recall the discussion re equation (7.1). But since course choice behaviour across colleges is mostly consistent with variation in the general measures, it is a reasonable presumption that any such biases as may exist are not noticeably affected by choice of school.)

Other tests, requiring development work at this stage, would abandon the shuffling of the general measures v_i and simulate the error variables instead. The simplest of these would test the one-factor model, but it is also appropriate to compare results from such a test with those coming from a two-factor model. This work would overlap with both data analyses investigating the two-factor model (3.14) and possible constructions of multiple aggregates considered briefly later in this chapter.

In view of the simulations run already and the theoretical analysis in Daley (1988), I would expect such simulations to confirm the thrust of the empirical analyses noted in the main body of this report.

The 1977–85 ACT Algorithm — Model 1

I start by introducing notation in the context of what appeared to me to be the existing ACT procedure when I first saw it in 1984; some of this view has been reflected in Daley (1985).

Scaled scores Y_{ij}, formed by linear transformation from the school-based raw scores X_{ij}, are expressible in terms of ASAT scores A_i, which represent student $i's$ ability *cum* achievement index, except for error terms e_{ij}, so that

$$a_j + b_j X_{ij} = Y_{ij} = A_i - e_{ij} \,, \tag{13.1}$$

errors being attributable to imprecisions in the scores X_{ij} (equivalently, Y_{ij}). Thus,

$$A_i = Y_{ij} + e_{ij} \,, \tag{13.2}$$

and within any given sub-population, notably the candidature $\mathcal{C}_j$ of course j, we should have the moments of $\{A_i : i \text{ in } \mathcal{C}_j\}$ and $\{Y_{ij} : i \text{ in } \mathcal{C}_j\}$ agreeing. Then if we also assume that the first moment of e_{ij} is approximately zero, and that the variance of $\{e_{ij} : i \text{ in } \mathcal{C}_j\}$ and its covariance with $\{Y_{ij} : i \text{ in } \mathcal{C}_j\}$ are such that

$$\operatorname{var}_j(e_{ij}) + 2\operatorname{cov}_j(e_{ij}, Y_{ij}) \approx 0, \tag{13.3}$$

we should have both

$$\operatorname{ave}_j(A_i) \approx \operatorname{ave}_j(Y_{ij}) = a_j + b_j X_{ij} \tag{13.4}$$

and

$$\operatorname{var}_j(A_i) \approx \operatorname{var}_j(Y_{ij}) = b_j^2 \operatorname{var}_j(X_{ij}). \tag{13.5}$$

For convenience, we shall call the equations (13.4) and (13.5) the *existing scaling parameter equations.* All we have done is to give equations (13.1)–(13.3) as the core of a model from which to derive the equations that are currently used. En route we have stated certain assumptions explicitly, to which we shall return shortly. We stress that,

> **if the data are inconsistent with consequences of these assumptions, then some substitute assumptions must be made in order to justify the continued use of these equations**.

All I am attempting to do here, is to present a scenario that appears to be consistent with the thinking as it emerges from reading between the lines, because to the best of my knowledge

> **there is no comprehensive published or written version of a set of assumptions that leads to these two sets of scaling equations**

TABLE 13.1

Summary Statistics in Simulation Study of ASAT Scaling

Subst'n #	corr(TE$_i$, A_i)	ave(D_i)	SD(D_i)	SD(TE$_i$/3.6)
Original	0.717			
1	0.812	3.89	22.38	23.50
2	0.841	1.58	11.89	23.56
3	0.870	1.04	11.15	24.07
4	0.880	0.49	5.52	24.04
5	0.895	0.63	7.23	24.45

— or if there is, it is not readily accessible.

Observe that equation (13.2) implies that the candidature $\mathcal{C}_j$ in course j could equally well have had the set of scores

$$\{A_i + \mu_i\} = \{Y_{ij} + \mu_i + e_{ij}\}$$

for any $\{\mu_i\}$ because this does not change the relationship between $\{Y_{ij}\}$ and $\{A_i\}$. A test of the model can thus be constructed by (random) interchange of the scores A_i coupled with the appropriate simultaneous shift of all Y_{ij}. Specifically,

$$\text{simultaneously replace } \; A_i \text{ by } A'_i \; \text{ and } \; Y_{ij} \text{ by } Y'_{ij} \equiv Y_{ij} + A'_i - A_i. \tag{13.6}$$

Then if the scaling procedure and model are consistent, since the changes have left unaltered both the error variables e_{ij} and the component of Y_{ij} correlated with the e_{ij}, and if moreover the changes are executed randomly in the sense of being otherwise independent of the scores, we should expect that (e.g.) TE scores, which are interpretable by (13.1) as

$$TE_i = 3.6A_i + \sum_{\text{best } 3.6 \text{ scores}} (-e_{ij}), \tag{13.7}$$

should likewise be altered simply to $3.6A'_i + (TE_i - 3.6A_i)$. In other words,

$$D_i \equiv (TE'_i - TE_i) - 3.6(A'_i - A_i) \tag{13.8}$$

is an estimator of zero. Similarly, from (13.5), the parameters b'_j satisfying (13.9)

$$\text{var}_j(A'_j) \approx \text{var}_j(Y'_{ij}) = (b'_j)^2 \, \text{var}_j(X'_{ij}) \tag{13.9}$$

should have $b'_j/b_j \approx 1$.

This substitution routine has been executed on one data set for which corr(T_i, A_i) $\approx$ 0.70. Table 13.1 summarizes the effect of 5 successive passes through the substitution operation, in which student i retains the same deviations e_{ij} in the same set of subjects at the time of making the changes in ASAT and course scores. (Specifically, suppose that after the n^{th} substitution the ASAT and scaled scores are A_i^n and Y_{ij}^n, so that the residual scores are $e_{ij}^n = A_i^n - Y_{ij}^n$; then replace A_i^n by A_i^{n+1} and rescale the scores $\{A_i^{n+1} + e_{ij}^n\}$, yielding $\{Y_{ij}^{n+1}\}$ and thus a new set of residual scores.)

Table 13.2

Some Location and Scale Parameters in the first 4 Substitutions underlying Table 13.1

Subst'n #	English Group		Mathematics Group	
	a_j	b_j	a_j	b_j
1	–0.084	1.0468	–0.182	1.1518
2	–0.010	1.0053	0.027	1.0137
3	–0.002	1.0011	0.014	1.0028
4	–0.001	1.0002	0.008	1.0010

The trends in trials 1–5 in Table 13.1 indicate rapid divergence of the data set from its original character. In comparison with the results of other trials below, it is not the substitution routine that does this but rather the nature of the model.

Conclusion 13.1. **The 1977–85 scaling procedure and the model described by equations (13.1)–(13.3) are not mutually consistent.**

For the record, the parameters (a_j, b_j) for the moderation groups with English and Mathematics scores relative to their values in the preceding substitution (or original set) are as in Table 13.2. These serve merely to emphasize that, no matter how the original data may deviate from the model, these deviations are largely lost as successive substitutions force the data to take on some character implicit in the substitutions.

The 1977–85 ACT Algorithm — Model 2

The major deficiency of the model of the preceding section lies in the assumption, implicit in (13.1) and succeeding relations, that ASAT scores have far smaller errors than course scores in the context of the model used to describe the data set consisting of both course and ASAT scores. Modify that assumption so as to regard ASAT scores, along with course scores, as being described by the one-factor model at (4.21), and instead of replacing the ASAT scores *per se*, use for each student the estimate $\tilde{v}_i$ of the general achievement factor to find the estimates of the residuals in each score, and substitute for each estimate of the achievement factor another such estimate: in order to maintain the same distribution of such factors, replace $\tilde{v}_i$ by $\tilde{v}_{i+1}$ for some enumeration of the population (and identify $\tilde{v}_{N+1}$ with $\tilde{v}_1$ in a college of N students). Recompute scaling parameters, and repeat the whole procedure a number of times. If the model holds, the moments of the estimators of the discrepancies D_i should be close to zero (cf. Table 13.1 where this is not so!).

Following this procedure through a full cycle of a college data set revealed a far slower rate of deviation than is evident in Table 13.2. Another way of following such changes is to consider the difference in TE scores after successive shuffles much as at (13.8) but with estimates of v_i replacing the ASAT scores there. The differences are no longer decreasing monotonically to zero, but change sign (the first few are -1.41, -0.73, 0.33, -0.02, -0.42, 0.42, 0.40, -0.48, -1.69,. . .), while their standard deviations are no longer monotonically decreasing either. A long term trend is evident, and to show this, Table 13.3 give the mean square of the differences in some sets of 10 consecutive trials. Finally, as a measure of both the stability and the slow changes, when the first cycle through the population is complete, the mean difference equals 0.16 (which is neither small nor large in comparison with other values), and the standard deviation is 1.25, which is smaller than generally occurs but not the smallest such statistic.

Table 13.3

Means and Standard Deviations of Differences of
Adjusted TE Scores after Shuffling Estimators of v_i

Trials	Mean square difference	SD of Diff'ce.
1–10	0.655	4.13
31–40	0.476	3.47
64–73	0.222	2.56
101–110	0.137	1.97
full cycle	0.16	1.16

Method-of-Moment Scaling

Another sequence of trials has been run in which instead of the existing scaling procedure, Method-of-Moment scaling is used to estimate the parameters, but with the same residuals used in each case. Thus, this sequence of trials always starts from the given residuals, and there are no trends to observe as each $\tilde{v}_i$-permutation is equivalent to any other.

Table 13.4 gives various data for the first 56 substitutions. The correlations of shuffled ASAT and TE scores are shown in the last column, showing no trend as just noted (cf. Tables 13.1 and 13.3). The standard deviation of differences of TE scores is of similar size to what is shown in Table 13.3, but if the two sets of TE scores concerned are given the same first two moments, this falls sharply (results not shown), indicating that the rankings implied by the two TE scores are much closer than appears without such standardization.

This set of Monte Carlo shuffling trials covers just one college, and is certainly not complete. Yet the indications from them are consistent with the conclusions drawn earlier, in e.g. Chapter 8 so far as precision is concerned, from Chapter 10 in terms of correlations between course scores and scaling criteria. For the one college whose data have been used, it appears to be the case that the Method-of-Moment procedure leads to the smallest changes as the estimates of $\{v_i\}$ are swapped amongst themselves and the pseudo data sets reconstructed and rescaled. This is consistent with Conclusion 8.2. It is of interest in such experiments to sort out the relative impact of ASAT scores on one hand and the Other Course Score scaling procedure facets on the other, so far as deviations of the data from the models of Chapter 3 are concerned. This is said from the standpoint stressed earlier that

> **any scaling procedure is "fair" only to the extent to which the data are consistent with the assumptions used to construct the procedure, and in this regard, the existing procedure is far from being as fair as is easily and reasonably attainable**.

Conclusion 13.2. **The major component of the data set requiring statistical scrutiny to ensure fair use of a statistical scaling procedure concerns the ASAT scores.**

Conclusion 13.3. **The existing scaling parameter equations (13.4)–(13.5) are not justified by Model 1 but are supported a one-factor model as in Model 2. Method-of-Moment estimators with Other Course Score scaling may give an even more consistent fit to the one-factor model description of the data.**

Table 13.4

Statistics from Shuffling Estimates of v_i with Constant Residuals using Other Course Score Scaling with Method-of-Moment Estimation

Multiple Aggregates

There are comments in Masters & Beswick (1986) and implications in some of *MATHEF*'s conclusions (see in particular §§6.39–44 and its Recommendations 2 and 3) that in principle it should be possible to produce for some students more than one aggregate score, each reflecting measures of achievement in different areas of study. In this section I discuss briefly the question of having two aggregates, which for convenience I call briefly Q- and V-aggregates.

In terms of the representations of Chapter 2 or the models of Chapter 3, each course score Y_{ij} can be represented either as a sum of principal components or else as

$$Y_{ij} = v_i + \gamma_j v_{i2} + e_{ij} \tag{13.10}$$

where the terms e_{ij} represent error, have mean zero, and are uncorrelated with the general achievement measure v_i and contrast factor v_{i2}. Assume that v_{i2} is positively correlated with the difference $Q_i - V_i$. Then course scores for which $\gamma_j > 0$ are those that reflect more highly developed Q-skills, while for $\gamma_j < 0$ the scores reflect more highly developed V-skills. An averaged aggregate of scores for which all $\gamma_j > 0$ estimates $v_i + C_i' v_{i2}$ (plus error that has mean zero) for some positive C_i', and may be called an averaged Q-aggregate. Similarly, an averaged aggregate of course scores for which $\gamma_j < 0$ estimates $v_i - C_i'' v_{i2}$ for positive C_i'', and is an averaged V-aggregate. In educational jargon, these two measures reflect relative achievement in the quantitative and verbal domains respectively.

If a course curriculum is defined externally as in Queensland or for a public examination system, the nature of an achievement measure in the course changes only slowly with time and so can be readily classified by its coefficient γ_j. Under the ACT system courses are defined within each college, and are given area codes common to the system as a whole, though the courses themselves can be constituted as a selection of several units and thus need not be anything like a common course of study for students even within a college. It is thus more problematic to classify courses *a priori* as contributing to a Q- or V-aggregate.

Nevertheless, supposing such a classification has been agreed, it follows from the discussion of Chapters 2 and 3 that an aggregate should be constructed for each group of course scores using a one-factor model and Method-of-Moment estimation for the parameters (assuming that a linear transformation is satisfactory). Likewise, with only school-based assessment as in the ACT and Queensland, a reference scale is needed to establish comparability of scores between colleges, and the principles illustrated by the discussion of Chapters 4 to 7 apply.

In practice, educational criteria would play a role in constructing classifications: this being the case, statistical criteria should be used at the very least in post mortem studies to verify that the educational arguments are meeting the goals they claim. Also, it may be necessary to impose curriculum constraints to ensure that students' course choices have a chance of meeting educationally desired goals.

Current scaling practice in South Australia crudely resembles the use of classifications as above to construct two aggregates as an intermediate step in the construction of a single aggregate that can be regarded as estimating $v_i + |C_i v_{i2}|$ for student i.

Another practical difficulty concerns the construction of the Q- and V-estimators from a sufficient number of scores to warrant any aggregate measure being regarded as a good approximation to a well-defined "signal" component as distinct from reflecting predominantly a substantial random component or "noise".

CONCLUSION 13.4. **If multiple aggregates are constructed within restricted subsets of courses, the principles of Other Course Score scaling using Method-of-Moment estimators, and construction and use of reference scales as in Chapters 4–7, apply.**

References

Aitkin, M. A. (1968). The ranking of candidates at an examination. In *Mathematics in the Social Sciences in Australia* (Australian UNESCO Seminar, May 1968), 531–547. Australian Government Publishing Service, Canberra, 1972.

ASAT and TE Scores (1988). *ASAT and TE Scores: A Focus on Gender Differences*. Queensland Board of Secondary School Studies, Brisbane.

Beswick, D., Schofield, H., Meek, L. & Masters, G. (1984). *Selective Admission Under Pressure*. Report to Commonwealth Tertiary Education Commission. Centre for the Study of Higher Education, University of Melbourne.

Breland, H. M. & Griswold, P. A. (1982). Use of a performance test as a criterion in a differential validity study. *J. Educ. Psych.* **74**, 713–721.

Cook, J. S. & Cooney, G. H. (1976). The scaling of marks at the Higher School Certificate examination. *Reflections* **1**, 16–43.

Cooney, G. H. (1975). Standardization procedures involving moderator variables — some theoretical considerations. *Aust. J. Educ.* **19**, 50–63.

— (1976a). Scaling procedures: a review of Australian practices. *Aust. Math. Teacher* **32**, 57–62.

— (1976b). Identification of the major observed skills measured by the NSW Higher School Certificate. *The Australian University*, 1976, 196–201.

— (1978). A critique of standardization by bivariate adjustment — a rejoinder. *Aust. J. Educ.* **19**, 323–325.

Daley, D. J. (1984). Producing consistent Tertiary Entrance Scores in the ACT. Dept. Statistics (IAS), Australian National University (mimeo). [Distributed by ACT Schools Accrediting Agency to secondary colleges, February, 1985.]

— (1985a). Standardization by bivariate adjustment of internal assessments: Sex bias and other statistical matters. *Aust. J. Educ.* **29**, 231–247.

— (1985b). How should NSW HSC examination marks be reported? *Independent Education* **15** (2), 34–38.

— (1986a). Different sex differences from different modes of assessment: common experiences in three countries. Manuscript, Statistics Dept. (IAS), Australian National University.

— (1986b). “Fair” selection for Tertiary admission. Paper delivered at 25th Jubilee Conference, Centre for Administrative and Higher Education Studies, Univ. New England, Armidale NSW, October 1986. [Reproduced in Daley, 1987b]

— (1987a). Ranking in a one-factor model used to describe exam. marks. *Proc. Internat. Workshop/Seminar on Statistical Inference Procedures in Ranking and Selection*, Sydney, August 1987. American Sciences Press, Ohio (to appear).

— (1987b). Scaling NSW HSC marks for school-leaver admission. Report to the Canberra College of Advanced Education.

— (1988). Modelling examination marks, II. Technical Report (mimeo Series #1745), Department of Statistics, University of North Carolina at Chapel Hill.

— (1989). The existence of sex bias in ACT Tertiary Entrance scores. (Sketch of manuscript.)

— & Eyland, E. A. (1987). The new and old HSC: Figures, facts and fantasies. *Independent Education* **17** (3), 22–25.

— & Seneta, E. (1986). Modelling examination marks. *Aust. J. Statist.* **28**, 143–153.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. **38**. Society for Industrial and Applied Mathematics, Philadelphia.

— & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**, 54–77.

Hasofer, A. M. (1977). Some mathematical problems in mark scaling. *Aust. Math. Soc. Gazette* **4**, 55–60.

— (1978). A critique of standardization by bivariate adjustment. *Aust. J. Educ.* 19, 319–322.

Jolliffe, I. T. (1986). *Principal Component Analysis.* Springer–Verlag, New York.

Keeves, J. P., McBryde, B. & Bennett, L. A. (1977). The validity of alternative methods of scaling school assessments at the HSC level for the colleges and high schools of the Australian Capital Territory. *AARE Conference Proceedings*, 1977, 1–13.

Kuder, G. F. & Richardson, M. W. (1937). The theory and estimation of test reliability. *Psychometrika* **2**, 151-160.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Addison–Wesley, Reading, Mass.

Manly, B. F. J. (1988). The comparison and scaling of student assessment marks in several subjects. *Appl. Statist.* **37**, 385–395.

Marcus, M. & Newman, M. (1961). The permanent of a symmetric matrix, Abstract 587–85. *Notices Amer. Math. Soc.* **8**, 595.

Masters, G. N. & Beswick, D. G. (1986). *The Construction of Tertiary Entrance Scores: Principles and Issues.* Report of commissioned study, Centre for the Study of Higher Education, University of Melbourne.

MATHEF (1986). *Making Admission to Higher Education Fairer.* Report of the Committee for the Review of Tertiary Entrance Score Calculations in the Australian Capital Territory. Australian Capital Territory Schools Authority, The Australian National University, and Canberra College of Advanced Education.

McGaw, B. (1977). The use of rescaled teacher assessments in the admission of students to tertiary study. *Aust. J. Educ.* **21**, 209–225.

— (1987). Selection of students for higher education. *Unicorn* **13** (1), 4–9.

Morgan, D. E. (1979). Sex bias analysis: ASAT and Tertiary Entrance scores. Research paper, 1978 ACT Year 12 Certificate Research, ACT Schools Accrediting Agency. (Distributed to ACT Secondary Colleges, March 1979).

Morgan, G. & McGaw, B. (1988). Proposed combinations of Australian Scholastic Aptitude Test subscores for scaling course scores in the ACT. Australian Council for Educatioonal Research, Preliminary Report, 10pp.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.

Seneta, E. (1981). *Non-negative Matrices and Markov Chains* (2nd Ed.). Springer–Verlag, New York.

— (1984). A technical note on the proposed University of Sydney scaling system. (Unpublished report, Dept. Mathematical Statistics, Univ. Sydney).

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.* **35**, 876–879.

Tertiary Entrance in Queensland: A Review (1987). Report of the Working Party on Tertiary Entrance. Board of Secondary School Studies, Brisbane.

Year 12 Study. Information Paper published annually by ACT Schools Authority, Canberra.

Reset version of 1988 typescript

Appendix 1

Modelling Examination Marks, II

D. J. DALEY

Statistics Department
University of North Carolina at Chapel Hill
(visiting)[29]

ABSTRACT

In a given population $\mathcal{C}$ suppose that student i studies a subset $\mathcal{S}_i$ of the subjects offered in a curriculum and that marks or assessment scores $\{X_{ij} : i \text{ in } \mathcal{C}_j\}$ are given reflecting the (relative) achievements of the candidature $\mathcal{C}_j$ of the students taking subject j. Various estimates of scale parameters $\{\beta_j\}$ in the one-factor model

$$\alpha_j + \beta_j X_{ij} \equiv Y_{ij} = v_i + e_{ij}$$

for uncorrelated error variables $\{e_{ij}\}$ are examined for unbiasedness: those based on a method of moments approach appear to be asymptotically optimal. Further, for the range of values of $\mathrm{var}(v_i)$ and $\mathrm{var}(e_{ij})$ encountered in practice, the same estimators are fairly robust against the two-factor model

$$Y_{ij} = v_{i1} + \gamma_j v_{i2} + e'_{ij}$$

in which verbal/quantitative contrast factor measures $\{v_{i2}\}$ supplement the general factors $\{v_{i1}\}$, while the sizes of the residuals $\{e'_{ij}\}$ are then close to the known sizes of the measurement errors they incorporate.

1. Introduction

This paper is a sequel to an earlier joint work (Daley & Seneta, 1986), referred to below as (I), in which a one-factor model was proposed to describe a data set $\{X_{ij}\}$ such as the examination marks obtained by students at the Year 12 level at which they complete their secondary schooling and seek entry to a tertiary institution on the basis of their exam. marks as their academic record. Within the several systems operating in Australasia, such data sets have for some one to two decades been the source for the prime or even sole determinant of entry to university or college of advanced education. The earlier paper and this are concerned with examining the basis for such determination which, for all that it has been accepted at large, is neither well understood nor administered with the degree of impartiality and sophistication that might be hoped for. This is said because the analyses that follow from the discussion below lead to the conclusion that, if the existing procedures were brought into line with what is being attempted, then for admission to some tertiary level courses, proportions of up to some 10 to 20 or even 30 per cent. of students who at present gain admission would be replaced by others. Unquestionably, existing procedures are technically sloppy; worse, the sloppiness exists to an extent that there are observable consequences

[29] Permanent address: Statistics Dept. (IAS), Australian National University, G.P.O. Box 4, Canberra A.C.T. 2601, Australia.

of appreciable size. However, it would be wrong to attach blame to some authorities, because the technical sloppiness has first to be exposed, and that in part is what this paper is about.

The data set $\{X_{ij}\}$ need not necessarily consist simply of examination marks. For example, the marks or scores may arise from school-based assessments, or from a combination of them with exam. marks, or from reference test scores such as Aptitude tests (e.g. SAT scores in USA or Australian Scholastic Aptitude Test scores). It will be convenient to call all such measures *scores* or *marks*, and to call the "subject" or "course" area from which they are derived a *subject* or *course*, even though there may not necessarily be a uniform and precisely defined "subject" for the individuals i given scores X_{ij} in the subject or course j.

The dominant issues in this paper are the consequences of estimating parameters $(\alpha_j,\ \beta_j)$ in the linear transformation

$$Y_{ij} = \alpha_j + \beta_j X_{ij} \tag{1.1}$$

and the resultant properties of the average score

$$Y_{i\cdot} \equiv \mathrm{ave}_i(Y_{ij}) \tag{1.2}$$

(formal first and second moment operators like $\mathrm{ave}_i(\cdot)$ are defined around (1.3)–(1.7) below). In particular we note the different properties of $Y_{i\cdot}$ that arise from a variety of possible estimation procedures for $(\alpha_j,\ \beta_j)$ and from a variety of possible model assumptions for $\{Y_{ij}\}$.

The basic notation used here is consistent with that of Daley (1988). The set $\mathcal{C}$ of all individuals with scores $\{X_{ij}\}$ is called the *candidature*. In general it will be a *sub-candidature*

$$\mathcal{C}_j \equiv \{i : i \text{ has a score } X_{ij}\}$$

of individuals taking a particular course j, because usually each student i is expected to choose only a subset $\mathcal{S}_i$ from all the subjects $\mathcal{S}$ in the curriculum. Student i has $n_i \equiv \#(\mathcal{S}_i)$ course scores and $N_j \equiv \#(\mathcal{C}_j)$ students take course j. Formal moments are defined as in

$$\mathrm{ave}_j(X_{ij}) = \sum_{i \text{ in } \mathcal{C}_j} X_{ij}/N_j\,, \tag{1.3}$$

$$[\mathrm{s.d.}_j(X_{ij})]^2 = \mathrm{var}_j(X_{ij}) = \sum_{i \text{ in } \mathcal{C}_j} [X_{ij} - \mathrm{ave}_j(X_{ij})]^2/N_j\,, \tag{1.4}$$

$$\mathrm{ave}_{jk}(X_{ij}) = \sum_{i \text{ in } \mathcal{C}_{jk}} X_{ij}/N_{jk}\,, \tag{1.5}$$

$$\mathrm{cov}_{jk}(X_{ij}, X_{ik}) = \sum_{i \text{ in } \mathcal{C}_{jk}} [X_{ij} - \mathrm{ave}_{jk}(X_{ij})][X_{ik} - \mathrm{ave}_{kj}(X_{ik})]/N_{jk}\,, \tag{1.6}$$

where $\mathcal{C}_{jk} = \mathcal{C}_j \cap \mathcal{C}_k$, $N_{jk} = \#(\mathcal{C}_{jk})$, and

$$\mathrm{corr}_{jk}(X_{ij}, X_{ik}) = \frac{\mathrm{cov}_{jk}(X_{ij}, X_{ik})}{[\mathrm{s.d.}_{jk}(X_{ij})][\mathrm{s.d.}_{kj}(X_{ik})]}\,. \tag{1.7}$$

Under (1.1),

$$\mathrm{ave}_j(Y_{ij}) = \alpha_j + \beta_j\, \mathrm{ave}_j(X_{ij})\,, \tag{1.8}$$

$$\mathrm{var}_j(Y_{ij}) = \beta_j^2\, \mathrm{var}_j(X_{ij}). \tag{1.9}$$

A major aim here is to elucidate what is entailed in basing ranking decisions on statistics like (cf. (1.2))

$$Y_{i\cdot} = \sum_{j \text{ in } \mathcal{S}_i} Y_{ij}/n_i \tag{1.10}$$

or more generally, for some subset $\mathcal{S}_i''$ of $\mathcal{S}_i$ that may depend on $\{Y_{ij} : j \text{ in } \mathcal{S}_i\}$,

$$Y_{i\cdot}'' = \sum_{j \text{ in } \mathcal{S}_i''} Y_{ij}/\#(\mathcal{S}_i'') \tag{1.11}$$

Whether it is specifically stated or not, Australasian practice has reflected as an act of faith that, no matter what scaling procedure has been used, a representation of the form

$$Y_{ij} = v_i + e_{ij} \tag{1.12}$$

then holds for certain error terms e_{ij} with zero mean, and that this is a valid unbiased representation. If this is so, then

$$Y_{i\cdot}'' = v_i + e_i'' \tag{1.13}$$

for some error term e_i'' that does not necessarily have zero mean but does have smaller variance than (almost all) e_{ij}. Empirically, a representation such as (1.12) does hold as a crude first approximation, implying that, when a recipe such as at (1.11) is followed, the representation at (1.13) necessarily holds in this crude sense. Because of this implication, it follows that when the parameter v_i in (1.12) is replaced by an estimate, unbiasedness of that estimate will be a highly desirable property.

There is an important practical reason for considering *linear* transformations such as at (1.1), relating to an invariance property of rankings as at (1.10) and (1.11). It is simply this, that if $\{(\alpha_j,\ \beta_j),\ v_i\}$ as in (1.1) and (1.12) are replaced by

$$\{(\alpha_j', \beta_j'), v_i'\} \equiv \{(A + \alpha_j B, \beta_j B), A + Bv_i\} \tag{1.14}$$

for any real A and real positive B, so that in place of Y_{ij} we should have

$$Y_{ij}' = A + \alpha_j B + \beta_j B X_{ij} = A + BY_{ij}\,, \tag{1.15}$$

the ranking as follows from (1.10) or (1.11) is unchanged, and the implication that (1.13) holds is likewise unchanged.

2. One-factor model

Whether unknowingly or explicitly as reported for example in Daley (1987) and Seneta (1987), existing Australasian mark-scaling procedures are based on an assumption that the one-factor model as in (I) provides a satisfactory description of the scores $\{X_{ij}\}$ concerned. The linearly transformed scores Y_{ij} as at (1.1), or more generally the transforms $f_j(X_{ij})$ for some family of monotonic increasing functions $f_j(\cdot)$, are assumed to have the structure

$$Y_{ij} = v_i + e_{ij} \tag{2.1}$$

for some unknown common (achievement) measure v_i and residual error terms e_{ij} that may embody both model-fit and measurement errors, such that, when viewed as random variables (r.v.s), the set $\{e_{ij} : i \text{ in } \mathcal{C}_j\}$ has

$$\mathrm{E}(e_{ij}) = 0, \qquad \mathrm{Var}(e_{ij}) = \sigma_j^2\,, \tag{2.2}$$

and is independent of both $\{v_i\}$ and sets $\{e_{ik}\}$ for $k \neq j$. (The use of $\mathrm{Var}(\cdot)$ as distinct from e.g. $\mathrm{var}_j(\cdot)$ is deliberate.) From these assumptions it follows that

$$\mathrm{Ave}_j(Y_{ij}) = \mathrm{ave}_j(v_i), \tag{2.3}$$

$$\mathrm{Var}_j(Y_{ij}) = \mathrm{var}_j(v_i) + \sigma_j^2\,, \tag{2.4}$$

$$\mathrm{Cov}_{jk}(e_{ij}, e_{ik}) = 0, \tag{2.5}$$

$$\mathrm{E}(\mathrm{cov}_j(v_i, e_{ij})) = 0, \tag{2.6}$$

$$\mathrm{Cov}_{jk}(Y_{ij}, Y_{ik}) = \mathrm{var}_{jk}(v_i). \tag{2.7}$$

3. Estimation in the one-factor model

The concern of this section is with the following questions:

If random variables $\{X_{ij}\}$ are such that $\{Y_{ij}\} \equiv \{\alpha_j + \beta_j X_{ij}\}$ satisfy the assumptions of section 2, what estimators of $\{(\alpha_j,\, \beta_j),\, v_i\}$ might be used, and what are their properties?

3.1. Maximum likelihood estimation

Suppose additionally, only in this sub-section, that the r.v.s e_{ij} are independently and normally distributed like $N(0,\, \sigma_j^2)$ r.v.s. Then the likelihood of the data set $\{X_{ij}\}$ is well-defined by

$$L \equiv \prod_{j \text{ in } \mathcal{S}} \prod_{i \text{ in } \mathcal{C}_j} (\sigma_j \sqrt{2\pi})^{-1} \exp[-(\alpha_j + \beta_j X_{ij} - v_i)^2 / 2\sigma_j^2]. \tag{3.1}$$

Suppose that for one particular j we have $(\alpha_j, \beta_j) = (0, 1)$ and $v_i = X_{ij}$. Then the term $\exp[\cdot]$ in (3.1) is identically one for this j, irrespective of σ_j^2, and L is maximized by setting $\sigma_j^2 = 0$ and takes the value $L = \infty$.

In comparison with the three other estimators from (I), the assumption of normality in order to have an expression for L is unnecessarily strong. Also, as in (I), it is unreasonable to assume that $\sigma_j^2 = 0$ for any j. Accordingly, this approach will be considered no further here.

3.2. Least squares estimation

It is frequently the case that there are close connections between least squares and maximum likelihood estimators. Accordingly, in reverting to the general one-factor model assumptions as outlined earlier without any assumptions of normality, we start by seeking estimators via the minimization of

$$S^2 \equiv \sum_{j \text{ in } \mathcal{S}} \sum_{i \text{ in } \mathcal{C}_j} (Y_{ij} - v_i)^2 = \sum_{j \text{ in } \mathcal{S}} \sum_{i \text{ in } \mathcal{C}_j} (\alpha_j + \beta_j X_{ij} - v_i)^2. \tag{3.2}$$

Inspection shows that S^2 is minimized with the value 0 by setting $\alpha_j = \beta_j = v_i = 0$ for all j in $\mathcal{S}$ and i in $\mathcal{C}_j$. Now the only conceivable boundary condition involving these parameters concerns the positivity of β_j, so we can expect that any other minimum of S^2 will be a solution $\{(\bar{a}_j,\, \bar{b}_j),\, \bar{v}_i\}$ of the least squares normal equations

$$\bar{v}_i = \sum_{j \text{ in } \mathcal{S}_i} (\bar{a}_j + \bar{b}_j X_{ij})/n_i \equiv \mathrm{ave}_i(\bar{a}_j + \bar{b}_j X_{ij}) = \mathrm{ave}_i(\bar{Y}_{ij}), \tag{3.3}$$

$$\bar{a}_j + \bar{b}_j\, \mathrm{ave}_j(X_{ij}) = \mathrm{ave}_j(\bar{v}_i), \tag{3.4}$$

$$\bar{b}_j\, \mathrm{var}_j(X_{ij}) = \mathrm{cov}_j(\bar{v}_i, X_{ij}). \tag{3.5}$$

Given any non-trivial solution $\{(\bar{a}_j,\, \bar{b}_j),\ \bar{v}_i\}$, inspection shows that for any real A and real positive B, the set $\{(A+B\bar{a}_j,\, B\bar{b}_j),\, A+B\bar{v}_i\}$ is also a solution. The all-zero solution is included in this set at the point $(A,\, B) = (0,\, 0)$.

Suppose without loss of generality that the data set $\{X_{ij}\}$ is generated by the model

$$X_{ij} = v_i + e_{ij}, \tag{3.6}$$

and that the set $\{(\bar{a}_j,\, \bar{b}_j),\ \bar{v}_i\}$ satisfies the equations (3.3)–(3.5) with $(\bar{a}_s,\, \bar{b}_s) = (0,\, 1)$ for some s (this last condition simply determines values of the parameters A, B within the class of solutions of the equations). Observe that the model quantities have $(\alpha_j,\, \beta_j) = (0,\, 1)$ for all j: we shall show that in general we cannot expect the least squares estimators to have these values as their long-run average values, and thereby conclude that

in general, the least squares estimators $\{(\bar{a}_j\,, \bar{b}_j),\ \bar{v}_i\}$ are biased. (3.7)

Start by assuming that all N_j are sufficiently large for the strong law of large numbers to hold, so that

$$\text{ave}_j(X_{ij}) = \text{ave}_j(v_i) + O_p(N^{-\frac{1}{2}}), \tag{3.8}$$

$$\text{var}_j(X_{ij}) = \text{var}_j(v_i) + \sigma_j^2(1 + O_p(N^{-\frac{1}{2}})). \tag{3.9}$$

Then, correct to terms that are $o_p(1)$ in $N.$,

$$\begin{aligned}\bar{v}_i = \text{ave}_i(\bar{a}_j + \bar{b}_j X_{ij}) = \text{ave}_i(\bar{a}_j + \bar{b}_j(v_i + e_{ij})) &= \text{ave}_i(\bar{a}_j) + v_i\,\text{ave}_i(\bar{b}_j) + \text{ave}_i(\bar{b}_j e_{ij}),\\ \text{Var}(\bar{v}_i) &= o(1) + o(1) + \text{ave}_i(\bar{b}_j^2\sigma_j^2),\\ \text{Var}_j(\bar{v}_i) &\approx o(1) + [\text{ave}_j(\text{ave}_i(\bar{b}_k))]^2\,\text{var}_j(v_i) + \text{ave}_j[\text{ave}_i(\bar{b}_k^2\sigma_k^2)]\,.\end{aligned}$$

In practice, there is interest in finding transformations that satisfy one of the two sets of constraints:

[1] $\{X_{is}\} = \{Y_{is}\}$ for some particular subject s, equivalently, $(\alpha_s,\, \beta_s) = (0,\, 1)$;

[2] $\sum_j \sum_i (X_{ij} - Y_{ij}) = 0 = \sum_j \sum_i (X_{ij}^2 - Y_{ij}^2)$, equivalently,

$$X_{\cdot\cdot} \equiv \text{ave}_{\text{all}}(X_{ij}) \equiv \frac{\sum_{j \text{ in } \mathcal{S}} N_j\,\text{ave}_j(X_{ij})}{\sum_{j \text{ in } \mathcal{S}} N_j} = \frac{\sum_{j \text{ in } \mathcal{S}} N_j\,\text{ave}_j(Y_{ij})}{\sum_{j \text{ in } \mathcal{S}} N_j} \equiv Y_{\cdot\cdot} \tag{3.10}$$

and

$$\begin{aligned}\text{var}_{\text{all}}(X_{ij}) \equiv \frac{\sum_j \sum_i (X_{ij} - X_{\cdot\cdot})^2}{\sum_{j \text{ in } \mathcal{S}} N_j} &= \frac{\sum_{j \text{ in } \mathcal{S}} N_j[\text{var}_j(X_{ij}) + (\text{ave}_j(X_{ij}) - X_{\cdot\cdot})^2]}{\sum_{j \text{ in } \mathcal{S}} N_j}\\ &= \frac{\sum_{j \text{ in } \mathcal{S}} N_j[\text{var}_j(Y_{ij}) + (\text{ave}_j(Y_{ij}) - Y_{\cdot\cdot})^2]}{\sum_{j \text{ in } \mathcal{S}} N_j} \equiv \text{var}_{\text{all}}(Y_{ij}).\end{aligned} \tag{3.11}$$

Either of these sets of constraints leads to seeking solutions $\{(\bar{a}_j, \bar{b}_j),\ \bar{v}_i\}$ of a modified set of equations. In the case of [1] for example, we should seek to minimize

$$S^2 + \lambda_1\alpha_s + \lambda_2(\beta_s - 1) \tag{3.12}$$

for Lagrangian multipliers λ_1, λ_2, leading to equations (3.3)–(3.5) as above for $j \neq s$ while for $j = s$ we have instead

$$\begin{aligned} \bar{a}_s + \bar{b}_s \operatorname{ave}_s(X_{is}) + \lambda_1 &= \operatorname{ave}_s(\bar{v}_i), \\ \bar{b}_s \operatorname{var}_s(X_{is}) + \lambda_2 &= \operatorname{cov}_s(\bar{v}_i, X_{is}). \end{aligned}$$

In view of the assumed identity $(\alpha_s, \, \beta_s) = (0, \, 1)$, these yield

$$\lambda_1 = \operatorname{ave}_s(\bar{v}_i) - \operatorname{ave}_s(X_{is}), \tag{3.13}$$
$$\lambda_2 = \operatorname{cov}_s(\bar{v}_i, X_{is}) - \operatorname{var}_s(X_{is}), \tag{3.14}$$

which with the other equations at (3.3)–(3.5) yield a set of $2\#(\mathcal{S}) + \#(\mathcal{C})$ linear equations in as many unknowns. In what follows, it is assumed that this set has a unique solution (cf. (I)). Observe that the convergence of any iterative routine for solving the equations is *prima facie* evidence for the existence though not necessarily the uniqueness of a solution.

The identities that follow from a similar treatment of the constraints at [2] are more suggestive because they can be written (with $X_{\cdot j} = \operatorname{ave}_j(X_{ij})$) in the forms

$$\sum_j N_j[\alpha_j + (\beta_j - 1)X_{\cdot j}] = 0, \tag{3.15}$$
$$\sum_j N_j[(\beta_j^2 - 1) \operatorname{var}_j(X_{ij}) + (\alpha_j + \beta_j X_{\cdot j})^2 - X_{\cdot j}^2] = 0, \tag{3.16}$$

from which we may anticipate that if the scores X_{ij} satisfy (3.6) with $\operatorname{ave}_{\text{all}}(v_i) = 0$, then when $n_i \approx n$ independent of i and v_i, both

$$\sum_j N_j(\beta_j - 1) \approx 0 \quad \text{and} \quad \sum_j (\beta_j - 1) \approx 0. \tag{3.17}$$

The expression to be minimized, with Lagrangian parameters λ_1 and λ_2, is now

$$S^2 + 2\lambda_1 \sum_j N_j[\alpha_j + (\beta_j - 1)X_{\cdot j}] + \lambda_2 \sum_j N_j[(\beta_j^2 - 1) \operatorname{var}_j(X_{ij}) + (\alpha_j + \beta_j X_{\cdot j})^2 - X_{\cdot j}^2].$$

The resulting normal equations can be written in the form

$$[\bar{a}_j + \bar{b}_j X_{\cdot j} - \operatorname{ave}_j(\bar{v}_i)] + \lambda_1 + \lambda_2[\bar{a}_j + \bar{b}_j X_{\cdot j}] = 0, \tag{3.18}$$
$$\sum_{i \text{ in } \mathcal{C}_j} X_{ij}[\bar{a}_j + \bar{b}_j X_{ij} - \bar{v}_i] + \lambda_1 N_j X_{\cdot j} + \lambda_2 N_j[\bar{b}_j \operatorname{var}_j(X_{ij}) + X_{\cdot j}(\bar{a}_j + \bar{b}_j X_{\cdot j})] = 0, \tag{3.19}$$

or equivalently, for each j in $\mathcal{S}$,

$$\lambda_1 + (1 + \lambda_2)(\bar{a}_j + \bar{b}_j X_{\cdot j}) = \operatorname{ave}_j(\bar{v}_i), \tag{3.18$'$}$$
$$(1 + \lambda_2)\bar{b}_j \operatorname{var}_j(X_{ij}) = \operatorname{cov}_j(\bar{v}_i, X_{ij}). \tag{3.19$'$}$$

Compare the Lagrangian multipliers here with the particular solution as below (3.5) with $A = \lambda_1$ and $B = 1 + \lambda_2$.

What now follows amplifies remarks[30] in (I) concerning the relative sizes of the least squares (LS) estimators $\bar{b}_j$ of β_j. Assume without loss of generality that the relations (3.6) hold and that LS estimators $\{(\bar{a}_j,\, \bar{b}_j),\, \bar{v}_i\}$ have been determined satisfying (3.3)–(3.5) subject to the constraints (3.10) and (3.11). What then is

$$\mathrm{E}(\bar{Y}_{ij}) = \mathrm{E}(\bar{a}_j + \bar{b}_j X_{ij}) = \mathrm{E}(\bar{a}_j) + \mathrm{E}(\bar{b}_j) v_i \text{ ?}$$

Observe that for large N_j,

$$\begin{aligned} \mathrm{var}_j(X_{ij}) &\approx \mathrm{var}_j(v_i) + \sigma_j^2 , \\ \mathrm{cov}_j(\bar{v}_i, X_{ij}) &= \mathrm{cov}_j \Big(\sum_{k \text{ in } \mathcal{S}_i} (\bar{a}_k + \bar{b}_k X_{ik})/n_i ,\ X_{ij} \Big) \\ &= \mathrm{cov}_j \Big(\sum_{k \text{ in } \mathcal{S}_i} (\bar{a}_k + \bar{b}_k v_i + \bar{b}_k e_{ik})/n_i ,\ v_i + e_{ij} \Big) \\ &\approx \mathrm{ave}_j \left(\mathrm{ave}_i(\bar{b}_k) \right) \mathrm{var}_j(v_i) + \bar{b}_j \sigma_j^2 / n \end{aligned}$$

where $1/n = \mathrm{ave}_{\mathrm{all}}(1/n_i)$. Consequently,

$$\bar{b}_j \approx \frac{\mathrm{ave}_j(\mathrm{ave}_i(\bar{b}_k)) + \bar{b}_j/(n\Gamma_j)}{(\lambda_2 + 1)(1 + 1/\Gamma_j)} \tag{3.20}$$

where

$$\Gamma_j = \mathrm{var}_j(v_i)/\sigma_j^2 \,. \tag{3.21}$$

In practice, $\mathrm{cov}_j(\bar{v}_i,\, X_{ij}) > 0$ so $1 + \lambda_2 > 0$, $n \approx 5$, and $\Gamma_j \approx 3$ to 6. Noting that $\mathrm{E}(\bar{b}_j) = 1$ for unbiased $\bar{b}_j$, we ask **how much does $\bar{b}_j$ then differ from 1**? Consider two scenarios. First, suppose the iterated average in (3.20) equals 1; this requires $1 + \lambda_2 \approx \frac{5}{6}$ and the range for $\bar{b}_j$, which is then a function of Γ_j, is about 0.96 to 1.07. For the second scenario, noting that it tends to be the case that students have more courses with Γ_k in common, and hence $\bar{b}_k$ in common, replace the double average by $\frac{1}{2}(1 + \bar{b}_j)$; using $1 + \lambda_2 \approx \frac{5}{6}$ again, $\bar{b}_j$ is now about 0.92 to 1.14. In either case, the assertion at (3.7) is supported, and it is supported more strongly by the example involving what appear to be the more realistic approximations.

3.3. Estimation via mean and variance equating

Suppose given just two sets of scores, $\{X_{i0}\}$ and $\{X_{i1}\}$ say, and suppose that the latter set is such that, after some unknown linear transformation as at (1.1), the resulting scores $\{Y_{i1}\}$ have the same structure (3.6) as $\{X_{i0}\}$ with $\mathrm{Var}(X_{i0} - v_i) = \mathrm{Var}(Y_{i1} - v_i)$ for all i. In this bivariate context, the following estimation procedure is asymptotically appropriate, remaining so in the multivariate context *provided* that the variance terms σ_j^2 are constant for all j, a condition which is not met in practice.

Notwithstanding the absence of such justification in terms of consistency with any model, it has been common practice to use as estimators of $\{(\alpha_j,\, \beta_j),\, v_i\}$ the attempted "solution" to the set of equations

$$\bar{\bar{v}}_i = \mathrm{ave}_i(\bar{\bar{a}}_j + \bar{\bar{b}}_j X_{ij}), \tag{3.22}$$

$$\bar{\bar{b}}_j = \left(\frac{\mathrm{var}_j(\bar{\bar{v}}_i)}{\mathrm{var}_j(X_{ij})} \right)^{1/2} = \frac{\mathrm{s.d.}_j(\bar{\bar{v}}_j)}{\mathrm{s.d.}_j(X_{ij})} \,, \tag{3.23}$$

$$\bar{\bar{a}}_j + \bar{\bar{b}}_j X_{\cdot j} = \mathrm{ave}_j(\bar{\bar{v}}_i). \tag{3.24}$$

[30] Masters and Beswick (1986), in quoting remarks from (I) about least squares estimators at their §2.49, erroneously inferred that they apply to method-of-moment estimators.

Because the ratio at (3.23) < 1, both in theory and in practice, these equations when iterated converge to the degenerate (null) solution. This inconsistency has been resolved in practice by fixing the estimated scale parameters $\bar{\bar{b}}_j$ after one or two iterations and determining $\bar{\bar{a}}_j$ for such fixed $\bar{\bar{b}}_j$. In view of the invariance properties around (1.14) and (3.5), an alternative is to impose one of the sets of conditions at [1] and [2] above.

Provided now that both σ_j^2 and $\mathrm{var}_j(v_i)$ are independent of j, the procedure is consistent with the model as outlined. To see this, observe as earlier that if any consistent non-degenerate solution of these equations exists, then a family of such solutions will exist consistent with the invariance property already noted.

Next, suppose as earlier that the data set is generated as at (3.6). Then for

$$\mathrm{E}(\bar{\bar{Y}}_{ij}) \equiv \mathrm{E}(\bar{\bar{a}}_j) + \mathrm{E}(\bar{\bar{b}}_j)v_i\,,$$

it again suffices to consider just the scale parameter estimator. For large N_j, much as in Section 3.2, (3.9) holds while

$$\begin{aligned}\mathrm{var}_j(\bar{\bar{v}}_i) = \mathrm{var}_j\Big(\sum_{k\ \text{in}\ \mathcal{S}_i}(\bar{\bar{a}}_k + \bar{\bar{b}}_k X_{ik})/n_i\Big) &= \mathrm{var}_j\Big(\sum_{k\ \text{in}\ \mathcal{S}_i}(\bar{\bar{a}}_k + \bar{\bar{b}}_k v_i + \bar{\bar{b}}_k e_{ik})/n_i\Big)\\ &\approx \mathrm{ave}_j\{[\mathrm{ave}_i(\bar{\bar{b}}_k)]^2\}\,\mathrm{var}_j(v_i) + \mathrm{ave}_j[\mathrm{ave}_i(\bar{\bar{b}}_k^2\sigma_k^2)]/n\,.\end{aligned}$$

Consequently,

$$\bar{\bar{b}}_j^2 \approx \frac{\mathrm{ave}_j\{[\mathrm{ave}_i(\bar{\bar{b}}_k)]^2\}[1 + 1/(n\Gamma_j)]}{1 + 1/\Gamma_j} \tag{3.25}$$

and, as in the previous section, $\bar{\bar{b}}_j$ again varies with Γ_j, and in general is asymptotically biased on each side of 1. However, the bias is about half that of using LS estimators, so this method is preferable to LS estimation.

3.4. Method of Moments estimation

Refer back to the equations (2.2)–(2.7) where it was noted that, when (2.1) holds,

$$\mathrm{E}(\alpha_j + \beta_j\,\mathrm{ave}_j(X_{ij})) = \mathrm{ave}_j(v_i), \tag{3.26}$$

$$\mathrm{E}(\beta_j^2\,\mathrm{var}_j(X_{ij})) = \mathrm{var}_j(v_i) + \sigma_j^2\,, \tag{3.27}$$

$$\mathrm{E}(\mathrm{ave}_i(\alpha_k + \beta_k X_{ik})) = v_i\,. \tag{3.28}$$

Furthermore, from $v_i(\alpha_j + \beta_j X_{ij}) = v_i^2 + v_i e_{ij}$ we have

$$\mathrm{E}(\beta_j\,\mathrm{cov}_j(v_i, X_{ij})) = \mathrm{var}_j(v_i). \tag{3.29}$$

Method of moment estimation entails replacing the unknown parameters in (3.26)–(3.29) by their estimators $\{(\widetilde{a}_j, \widetilde{b}_j), \widetilde{v}_i\}$ say, and solving for them. Again, since as earlier the equations are no longer linear in the unknowns, an iterative solution scheme is adopted, of which a more extended account with a particular data set has been detailed in Daley (1987). For this, it is not necessary to use (3.27) other than to estimate σ_j^2 after finding all the other parameters, that is, *the method of moment estimators* $\{(\widetilde{a}_j, \widetilde{b}_j), \widetilde{v}_i\}$ *satisfy the equations*

$$\widetilde{a}_j + \widetilde{b}_j\,\mathrm{ave}_j(X_{ij}) = \mathrm{ave}_j(\widetilde{v}_i), \tag{3.30}$$

$$\widetilde{b}_j\,\mathrm{cov}_j(\widetilde{v}_i,\, X_{ij}) = \mathrm{var}_j(\widetilde{v}_i), \tag{3.31}$$

$$\widetilde{v}_i = \mathrm{ave}_i(\widetilde{a}_k + \widetilde{b}_k X_{ik}); \tag{3.32}$$

the (biased) estimator $\widetilde{s_j^2}$ of σ_j^2 is then given by

$$\widetilde{s_j^2} = \widetilde{b}_j^2 \operatorname{var}_j(X_{ij}) - \operatorname{var}_j(\widetilde{v}_i). \tag{3.33}$$

In practice, initial estimates such as $(\alpha_j,\, \beta_j) = (0,\, 1)$ are taken and successively iterated through (3.32), (3.31), and (3.30), as a perturbation analysis shows that under the conditions usually encountered, such a scheme then has satisfactory convergence properties.

To study the bias properties of the estimators of $\{\beta_j\}$ it again entails no loss of generality to assume that (3.6) holds. Then, asymptotically as before,

$$\begin{aligned}
\operatorname{var}_j(\widetilde{v}_i) &= \operatorname{var}_j\Big(\sum_{k \text{ in } \mathcal{S}_i} \frac{\widetilde{a}_k + \widetilde{b}_k X_{ik}}{n_i}\Big) = \operatorname{var}_j\Big(\sum_{k \text{ in } \mathcal{S}_i} \frac{\widetilde{a}_k + \widetilde{b}_k v_i + \widetilde{b}_k e_{ik}}{n_i}\Big) \\
&\approx \operatorname{ave}_j\{[\operatorname{ave}_i(\widetilde{b}_k)]^2\} \operatorname{var}_j(v_i) + \frac{\operatorname{ave}_j[\operatorname{ave}_i(\widetilde{b}_k^2 \sigma_k^2)]}{n}\,, \qquad (3.34) \\
\operatorname{cov}_j(\widetilde{v}_i, X_{ij}) &= \operatorname{cov}_j\Big(\sum_{k \text{ in } \mathcal{S}_i} \frac{\widetilde{a}_k + \widetilde{b}_k X_{ik}}{n_i}\,,\ X_{ij}\Big) = \operatorname{cov}_j\Big(\sum_{k \text{ in } \mathcal{S}_i} \frac{\widetilde{a}_k + \widetilde{b}_k v_i + \widetilde{b}_k e_{ik}}{n_i}\,,\ v_i + e_{ij}\Big) \\
&= \operatorname{ave}_j[\operatorname{ave}_i(\widetilde{b}_k)] \operatorname{var}_j(v_i) + \frac{\widetilde{b}_j \sigma_j^2}{n}\,. \qquad (3.35)
\end{aligned}$$

Thus,

$$\widetilde{b}_j \approx \frac{\operatorname{ave}_j[\operatorname{ave}_i(\widetilde{b}_k)]^2 \operatorname{var}_j(v_i) + \operatorname{ave}_j[\operatorname{ave}_i(\widetilde{b}_k^2 \sigma_k^2)]/n}{\operatorname{ave}_j[\operatorname{ave}_i(\widetilde{b}_k)] \operatorname{var}_j(v_i) + \widetilde{b}_j \sigma_j^2/n}\,, \tag{3.36}$$

which is closer to being unbiased than either of the two previous estimators. In particular, it is much less affected by variability of Γ_j. Note also that

$$\begin{aligned}
\operatorname{Var}(\widetilde{v}_i) &= \operatorname{Var}\Big(\sum_{k \text{ in } \mathcal{S}_i} \frac{\widetilde{a}_k + \widetilde{b}_k v_i + \widetilde{b}_k e_{ik}}{n_i}\Big) \\
&\approx \operatorname{Var}\Big(\sum_{k \text{ in } \mathcal{S}_i} \frac{\widetilde{a}_k + \widetilde{b}_k v_i}{n_i}\Big) + \sum_{k \text{ in } \mathcal{S}_i} \frac{\widetilde{b}_k^2 \sigma_k^2}{n_i^2} \\
&= O(N^{-1}) + n_i^{-1} \operatorname{ave}_i(\widetilde{b}_k^2 \sigma_k^2)\,. \qquad (3.37)
\end{aligned}$$

3.5. External reference measure

Suppose finally that a further set of scores $\{V_i\}$ is provided as estimators of $\{v_i\}$, so that

$$V_i = v_i + e_{iV} \tag{3.38}$$

where $\mathrm{E}(e_{iV}) = 0$, $\mathrm{E}(e_{iV}^2) = \sigma_V^2$, $\operatorname{cov}(v_i,\, e_{iV}) = 0$. It has been common in the educational measurement literature to use $\{V_i\}$ for "reference score equating" by requiring that

$$\operatorname{ave}_j(Y_{ij}) = \operatorname{ave}_j(V_i), \qquad \operatorname{var}_j(Y_{ij}) = \operatorname{var}(V_i), \tag{3.39}$$

in spite of its being known that biased estimators of the scale parameter β_j then ensue (see e.g. Cooney (1974, 1977), Hasofer (1977), and Potthoff (1982)), largely as a result it would appear that a model such as (2.1) was not in view, and in particular there was no suggestion of an approach via the estimation of v_i. In view of the model assumptions, equations (3.39) are equivalent to assuming that $\sigma_j^2 = \sigma_V^2$ for each j concerned. This assumption is similar to that of the mean and variance equating estimation procedure already outlined. On the one hand, it recognizes that both sets of scores $\{X_{ij}\}$ and $\{V_i\}$ are subject to error (i.e., imprecise determination), whether coming from model-fitting or actual measurement or both. On the other hand, it assumes that these errors are of the same size for all scores, whether from courses or the reference test, when in practice these are known to vary considerably (cf. the range 3 to 6 for Γ_j; for evidence, see Daley (1985) and Daley & Eyland (1987)).

Observe also that the error variance of the estimator is now proportional to σ_V^2/N_j rather than $\text{ave}_j[\text{ave}_i(\sigma_k^2)]/n \approx \sigma_j^2/(nN_j)$, and the increases from both $\sigma_j^2 < \sigma_V^2$ and $1/n < 1$ introduce appreciable errors into the estimation of v_i unless N_j is large.

Of even more concern is that estimates of the location parameters α_j are now prone to bias within the error variables e_{iV}. Broadly speaking these can be regarded as cultural biases, as for example concerning ethnicity and gender with SAT scores in USA and of gender in both Australia and UK (*MATHEF* (1986) refers to a survey paper manuscript of Daley).

3.6. Which estimation procedure?

In terms of precision of estimates, it is unquestionably the case that any of the *other course score* procedures of sections 3.2 to 3.4 is preferable to the external reference measure procedure of section 3.5. This is a simple consequence of the fact that, if a parameter v_i is estimated by several measures, and a measurement error (or, errors in variables) model is appropriate, then information on the parameter is derived better from a reasonable combination of all the observations contributing more or less equally rather than relying on a single set of observations. Within this group of procedures, the criterion of unbiasedness of the scale parameters $\{\beta_j\}$, which is relevant in the tails of the distribution of $\{v_i\}$ though less critical than unbiasedness of the location parameters $\{\alpha_j\}$, means the method of moment procedure of section 3.4 is to be preferred.

It is possible in principle to investigate these methods via either or both of Monte Carlo methods and resampling procedures. The major practical problem associated with using the former is to construct a data set consistent with both the model and the pattern of courses $\mathcal{S}_i$ taken by students in relation to their general measures v_i. One solution is to use the estimates of both $\{v_i\}$ and $\{\sigma_j^2\}$ from a data set (e.g., as from the method of moment procedure), and replace the observed errors by simulated values $\{e'_{ij}\}$ which should then be reasonably independent. For the latter, jackknife estimates of $\text{Var}(v_i)$ for example may be appropriate through the use of a common set of subsamples for different estimation procedures.

The estimation procedures of sections 3.2 to 3.4 can also be used in conjunction with an external reference measure such as $\{V_i\}$ by regarding the latter as a set of scores from some course, as for example regarding it as the course s as under the constraint [1] above (3.10). Such a procedure was adopted in the analyses to which brief reference is made in Chapter 5 of *MATHEF* (1986).

The one-factor model and its associated estimation procedures can be used on subsets of courses when the latter are chosen by some external prescriptive criteria. For example, *ad hoc* analyses have been performed on classifying courses j in $\mathcal{S}$ as lying in either a humanities (verbal) domain or a science and mathematics (quantitative) domain, and a procedure similar to that of section 3.5 followed within each of the two resulting subsets, whereas what has been sketched in section 3.4 would be much more appropriate. Again, all that is being reflected here is a lack of understanding of the logical need for any algorithm to be governed by a mathematical model that describes the context of the information being processed by the algorithm in such a way that, ideally, the principles underlying the algorithm and its application to the model are mutually consistent, optimal, and consistent with the data.

4. One-factor model procedures used on a two-factor model

It has been assumed so far that the one-factor model is a satisfactory description of the data in the sense that the sets of residuals $\{e_{ij}\}$ are mutually uncorrelated. (While we stated an assumption of independence at (2.2), all we have used, except for the maximum likelihood procedure which we have rejected, is this zero correlation property.) Since a *ranking* is a one-dimensional concept and the parameters $\{v_i\}$ correspond in a general sense to a first principal component of the multivariate set $\{Y_{ij}\}$, i.e., to the dominant component, it is arguable that at this stage it is enough to check that the resulting error terms are uncorrelated.

In practice, the data sets are such that a second component is always observable, and a third is also observable when certain external reference measures are used. It is therefore proper to consider the one-factor model estimation procedures in relation to these more detailed models. In this section we consider the following two-factor model which corresponds to the practical observation that many students tend to be relatively better in one of the two areas defined by a preponderance of verbal skills for one and quantitative skills for the other. (In more colloquial terms, students tend to be better in either the humanities area or the science and mathematics area.) Suppose then that we retain (1.1) but that instead of (2.1) we have

$$Y_{ij} = v_{i1} + \gamma_j v_{i2} + e'_{ij} \tag{4.1}$$

for some family of constants $\{\gamma_j\}$, general achievement measures $\{v_{i1}\}$, contrast measures $\{v_{i2}\}$, and residual variables $\{e'_{ij}\}$, such that over their common sub-candidatures, $\{v_{i1}\}$ and $\{v_{i2}\}$ are mutually uncorrelated and uncorrelated also with $\{e'_{ij}\}$. It is immediately recognizable that, in addition to the indeterminate parameters A, B as at (1.14) and (1.15) for the model at (2.1), there is another indeterminacy in the model at (4.1) in that the quantities $\{C\gamma_j\}$ and $\{C^{-1}v_{i2}\}$ yield the same description of any data set.

We content ourselves for the time being with observing that if (4.1) holds and we form the estimator of $Y_{i\cdot}$ at (1.10) by

$$\begin{aligned}\sum_{j \text{ in } \mathcal{S}_i} Y_{ij}/n_i &= \sum_{j \text{ in } \mathcal{S}_i} (v_{i1} + \gamma_j v_{i2} + e'_{ij})/n_i \\ &= v_{i1} + \frac{\sum_{j \text{ in } \mathcal{S}_i} \gamma_j}{n_i} v_{i2} + \frac{\sum_{j \text{ in } \mathcal{S}_i} e'_{ij}}{n_i} \\ &\equiv v_{i1} + \gamma'_{i\cdot}|v_{i2}| + e'_{i\cdot} \quad \text{say,}\end{aligned} \tag{4.2}$$

then again the dominant component is v_{i1} but, typically, because a student if anything tends to have a majority of courses from the area of relative strength in terms of the contrast measure v_{i2}, this dominant component is increased by a fraction of the contrast measure. (It is tacitly being assumed here that the coefficients γ_j lie in the range $(-1, 1)$ or thereabouts, by appropriate choice of the arbitrary constant C.) The last statement means that, no matter what convention has been adopted with regard to the sign of v_{i2}, each student will tend to have a majority of courses for which γ_j has the same sign as v_{i2}, and thus, taking

$$\gamma'_{i\cdot} \equiv \Big| \sum_{j \text{ in } \mathcal{S}_i} \gamma_j/n_i \Big| = \big| \operatorname{ave}'_i(\gamma_j) \big|,$$

the second term on the right-hand side of (4.2) is (usually) positive as implied.

The representation (4.2) makes little sense until we have some idea of the magnitude of the quantities involved. Our experience with data from three Australian sources indicates that

$\mathrm{var}_{\mathrm{all}}(v_{i1}) : \mathrm{var}_{\mathrm{all}}(v_{i2}) \approx 4 : 1$ or larger, that $\gamma_i' \approx 0.2$ to 0.5 , and that $\mathrm{Var}(e_{i\cdot}') \approx \mathrm{var}_{\mathrm{all}}(v_{i1})/10$ or less, so that the measures $Y_{i\cdot}$ can certainly be regarded as providing a classification of the population into several subgroups if that is required.

Questions of misclassification rates have been canvassed in Daley (1988).

A major benefit of having the representation (4.2) is that it explains observed covariances $\mathrm{cov}_{jk}(Y_{ij},\, Y_{ik})$ better than the one-factor model (2.1). To show this, we must make some assumptions that approximate the participation rates of students in various courses. To this end, assume that every student takes courses $j = 1$ and 2, these common courses being one in each of the major areas (e.g. every student takes English and Mathematics), that each student then takes three further courses in his area of strength, and that $\gamma_{i\cdot}' \approx |1.0 - 1.0 + 3(0.5)|/5 = 0.3$. Suppose also that 60% of students are in the area of course 1 and 40% in the other. (We could equally use 50% in each: we choose otherwise in order to illustrate effects of imbalance, of which the first is that the representation at (4.2) must be modified by replacing $|v_{i2}|$ by the top 60% of v_{i2} for the area of course 1, and the top 40% of $-v_{i2}$ for the area of course 2.) We shall suppose that the raw scores X_{ij} have the representation at (4.2), much as we made the assumption about $\beta_j = 1$ at (3.6) in our study of biases of b_j in section 3, with $\gamma_1 = -\gamma_2 = 1$ for the sake of definiteness. We have

$$\begin{aligned}
\textstyle\sum_i v_{i1} = 0 = \sum_i v_{i2} &= \mathrm{ave}_1(X_{i1}) = \mathrm{ave}_2(X_{i2}),\\
\mathrm{var}_1(X_{i1}) &= \mathrm{var}_1(v_{i1}) + \mathrm{var}_1(v_{i2}) + \mathrm{Var}(e_{i1}'),\\
\mathrm{var}_2(X_{i2}) &= \mathrm{var}_1(v_{i1}) + \mathrm{var}_1(v_{i2}) + \mathrm{Var}(e_{i2}');
\end{aligned}$$

assuming for the sake of argument that the measures $\{v_{i2}\}$ have the distribution $N(0, s_2^2)$ and that the 60% group takes courses 3, 5, and another, and that the 40% group takes courses 4, 6, and another, we have also

$$\begin{aligned}
\mathrm{ave}_3(X_{i3}) &= \mathrm{ave}_3(v_{i1}) + \tfrac12\,\mathrm{ave}\{\text{top } 60\% \text{ of } v_{i2}\} + 0\\
&= \mathrm{ave}_3(v_{i1}) + \tfrac12(0.644 s_2),\\
\mathrm{ave}_4(X_{i4}) &= \mathrm{ave}_4(v_{i1}) + \tfrac12\,\mathrm{ave}\{\text{top } 40\% \text{ of } -v_{i2}\} + 0\\
&= \mathrm{ave}_3(v_{i1}) + \tfrac12(0.965 s_2).\\
\mathrm{var}_3(X_{i3}) &= \mathrm{var}_3(v_{i1}) + \tfrac14\,\mathrm{var}\{\text{top } 60\% \text{ of } v_{i2}\} + \mathrm{Var}(e_{i3}')\\
&= \mathrm{var}_3(v_{i1}) + \tfrac14(0.650 s_2)^2 + \mathrm{Var}(e_{i3}'),\\
\mathrm{var}_4(X_{i4}) &= \mathrm{var}_4(v_{i1}) + \tfrac14\,\mathrm{var}\{\text{top } 40\%\text{of } -v_{i2}\} + \mathrm{Var}(e_{i4}')\\
&= \mathrm{var}_4(v_{i1}) + \tfrac14(0.560 s_2)^2 + \mathrm{Var}(e_{i4}').\\
\mathrm{cov}_{12}(X_{i1},\, X_{i2}) &= \mathrm{cov}_{12}(v_{i1} + v_{i2} + e_{i1}',\; v_{i1} - v_{i2} + e_{i2}')\\
&= \mathrm{var}_1(v_{i1}) - \mathrm{var}_1(v_{i2}),\\
\mathrm{cov}_{13}(X_{i1},\, X_{i3}) &= \mathrm{cov}_{13}(v_{i1} + v_{i2} + e_{i1}',\; \{v_{i1} + \tfrac12 v_{i2} + e_{i3}' : \text{top } 60\% \text{ of } v_{i2}\})\\
&= \mathrm{var}_3(v_{i1}) + \tfrac12\,\mathrm{var}_3(v_{i2})\\
&= \mathrm{var}_3(v_{i1}) + \tfrac12(0.650 s_2)^2\,,\\
\mathrm{cov}_{24}(X_{i2},\, X_{i4}) &= \mathrm{var}_4(v_{i1}) + \tfrac12(0.560 s_2)^2,\\
\mathrm{cov}_{14}(X_{i1},\, X_{i4}) &= \mathrm{cov}_{14}(v_{i1} + v_{i2} + e_{i1}', \{v_{i1} - \tfrac12 v_{i2} + e_{i4}' : \text{top } 40\% \text{ of } -v_{i2}\})\\
&= \mathrm{var}_4(v_{i1}) - \tfrac12\,\mathrm{var}_4\{v_{i2} : \text{top } 40\% \text{ of } -v_{i2}\}\\
&= \mathrm{var}_4(v_{i1}) - \tfrac12(0.560 s_2)^2\,,\\
\mathrm{cov}_{23}(X_{i2},\, X_{i3}) &= \mathrm{var}_3(v_{i1}) - \tfrac12(0.650 s_2)^2\,,\\
\mathrm{cov}_{35}(X_{i3},\, X_{i5}) &= \mathrm{var}_3(v_{i1}) + \tfrac14(0.650 s_2)^2\,,\\
\mathrm{cov}_{46}(X_{i4},\, X_{i6}) &= \mathrm{var}_4(v_{i1}) + \tfrac14(0.560 s_2)^2\,.
\end{aligned}$$

To investigate the effect of using the method of moment estimation procedure on the data as though they conform to the one-factor model, consider the result of calculation after the first iterative step:

$$\begin{aligned}
\mathrm{ave}_1(X_{i\cdot}) &= \mathrm{ave}_1(v_{i1}) + (0.3) \times (0.773 s_2) \\
&= 0.232 s_2 = \mathrm{ave}_2(X_{i\cdot}), \\
\mathrm{var}_1(X_{i\cdot}) &= \mathrm{var}_1(v_{i1}) + (0.3)^2 \times \mathrm{var}\{\text{top } 60\% \text{ of } v_{i2} \text{ and top } 40\% \text{ of } -v_{i2}\} + \mathrm{Var}(e'_{i\cdot}) \\
&= \mathrm{var}_1(v_{i1}) + (0.3)^2 \times (0.403 s_2)^2 + \mathrm{Var}(e'_{i\cdot}) = \mathrm{var}_2(X_{i\cdot}), \\
\mathrm{ave}_3(X_{i\cdot}) &= \mathrm{ave}_3(v_{i1}) + (0.3) \times (0.644 s_2), \\
\mathrm{var}_3(X_{i\cdot}) &= \mathrm{var}_3(v_{i1}) + (0.3)^2 \times (0.650 s_2)^2 + \mathrm{Var}(e'_{i\cdot}), \\
\mathrm{ave}_4(X_{i\cdot}) &= \mathrm{ave}_4(v_{i1}) + (0.3) \times (0.965 s_2), \\
\mathrm{var}_4(X_{i\cdot}) &= \mathrm{var}_4(v_{i1}) + (0.3)^2 \times (0.560 s_2)^2 + \mathrm{Var}(e'_{i\cdot}), \\
\mathrm{cov}(X_{i\cdot}, X_{i1}) &= \mathrm{var}_1(v_{i1}) + (0.3) \times \mathrm{cov}_1(\{\text{top } 60\% \text{ of } v_{i2} \text{ and top } 40\% \text{ of } -v_{i2}\}, v_{i2}) + \tfrac{1}{5}\mathrm{Var}(e'_{i1}) \\
&= \mathrm{var}_1(v_{i1}) + (0.3) \times (0.004 s_2^2) + (0.2) \times \mathrm{Var}(e'_{i1}), \\
\mathrm{cov}(X_{i\cdot}, X_{i2}) &= \mathrm{var}_1(v_{i1}) - (0.3) \times (0.004 s_2^2) + (0.2) \times \mathrm{Var}(e'_{i2}), \\
\mathrm{cov}(X_{i\cdot}, X_{i3}) &= \mathrm{var}_3(v_{i1}) + (0.15) \times \mathrm{cov}_3(v_{i2}, \{\text{top } 60\% \text{ of } \ v_{i2}\}) + (0.2) \times \mathrm{Var}(e'_{i3}) \\
&= \mathrm{var}_3(v_{i1}) + (0.15) \times (0.650 s_2)^2 + (0.2) \times \mathrm{Var}(e'_{i3}), \\
\mathrm{cov}(X_{i\cdot}, X_{i4}) &= \mathrm{var}_4(v_{i1}) + (0.15) \times (0.560 s_2)^2 + (0.2) \times \mathrm{Var}(e'_{i4}).
\end{aligned}$$

As the first iteration approximation to b_1 we have the ratio

$$\frac{\mathrm{var}_1(v_{i1}) + 0.0361 s_2^2 + \mathrm{Var}(e'_1)}{\mathrm{var}_1(v_{i1}) + 0.001 s_2^2 + (0.2)\mathrm{Var}(e'_{i1})]} .$$

Assuming that $\mathrm{Var}(e'_1) \approx \frac{1}{5}\mathrm{Var}(e'_{i1}) \approx \frac{1}{5} s_2^2$ and that $s_2^2 / \mathrm{var}_1(v_{i1}) \approx \frac{1}{5}$, this ratio ≈ 1.007, indicating that the effect of assuming that (4.2) holds with X_{ij} rather than Y_{ij} introduces a bias that is smaller than any of the biases considered in connection with estimates of β_j in section 3. Making similar assumptions in connection with the other courses leads to ratios that are likewise within 1% of 1.00. **Within the simplified course choices and using the typical values of variances for the ratios just considered, the method of moment estimation procedure derived from a one-factor model produces estimators for the two-factor model that are somewhat smaller than the bias terms canvassed for other course score estimation procedures in section 3. Accordingly, on these theoretical grounds, the one-factor model constructed via method of moment estimation produces adequate estimators even for the two-factor model as above.**

5. Reference test factor

It has long been known (Anastasi (1958) wrote of studies going back as far as 1929) that the relative performance of mid-teenage boys and girls on standardized tests such as SAT tests in USA differs from their relative performance under class-room assessment practices. Such differences would appear to be culturally based, or if not, a gender-linked interaction of the psyche with the mode of assessment in that multiple choice tests are predominantly used in standardized testing but not in most class-room based assessments. The presence of any such interaction is presumably not a gender trait per se but merely a gender-linked trait, in which case, if there exist at least two sets of pairs of assessments that can be regarded as being in similar areas, one from a standardized test or other multiple choice based test and the other from the classroom, then it should be possible to discern whether over all individuals it is feasible to postulate an analogue of (4.1) in the form

$$Y_{ij}(\delta_j) = v_{i1} + \gamma_j v_{i2} + \delta_j \Delta_i + e''_{ij} \tag{5.1}$$

where $\delta_j = +1$ or -1 and Δ_i denotes the relative performance of individual i as measured under two modes of assessment in course j.

Equation (5.1) has the consequence that if for example courses $j = 1$ and 2 have $\gamma_j = +1$ and -1 respectively, then the four sets of scores $\{Y_{i1}(1)\}$, $\{Y_{i1}(-1)\}$, $\{Y_{i2}(1)\}$, $\{Y_{i2}(-1)\}$, yield

$$\tfrac{1}{4}[Y_{i1}(1) + Y_{i2}(1) + Y_{i1}(-1) + Y_{i2}(-1)] = v_{i1} + e''_i(1), \tag{5.2}$$
$$\tfrac{1}{4}[Y_{i1}(1) - Y_{i2}(1) + Y_{i1}(-1) - Y_{i2}(-1)] = v_{i2} + e''_i(2), \tag{5.3}$$
$$\tfrac{1}{4}[Y_{i1}(1) + Y_{i2}(1) - Y_{i1}(-1) - Y_{i2}(-1)] = \Delta_i + e''_i(3), \tag{5.4}$$
$$\tfrac{1}{4}[Y_{i1}(1) - Y_{i2}(1) - Y_{i1}(-1) + Y_{i2}(-1)] = e''_i(4), \tag{5.5}$$

where on the assumption that the errors $e''_i(\cdot)$ are uncorrelated r.v.s with variances $\sigma_1^2, \ldots, \sigma_4^2$ say, the sets of error terms $\{e''_i(r) : i = 1, \ldots, N; r = 1, \ldots, 4\}$ are mutually uncorrelated with a common variance $\frac{1}{4}\sigma^2 \equiv (\sigma_1^2 + \cdots + \sigma_4^2)/16$. It is clear that some test of the model (5.1) is effected by forming the four linear contrasts (5.2)–(5.5) and finding their sums of squares, for which the respective expectations are

$$\mathrm{var}_1(v_{i1}) + \tfrac{1}{4}\sigma^2, \quad s_2^2 + \tfrac{1}{4}\sigma^2, \quad \mathrm{var}_1(\Delta_i) + \tfrac{1}{4}\sigma^2, \quad \text{and} \quad \tfrac{1}{4}\sigma^2. \tag{5.6}$$

Comparison of the observed mean squares with these expected mean squares, and in particular, that the last mean square is significantly smaller than any of the others, is evidence that (5.1) holds. Another test is effected by looking at the correlations of the sets of right-hand sides: near-zero correlations constitute additional evidence that (5.1) holds, being independent of the mean square evidence.

What is almost universally reported is that boys and girls differ in their relative abilities in the quantitative and verbal skill areas. In a report that admitted to having been written hastily, Masters and Beswick (1986) suggested and attempted to supply evidence that the gender-linked difference noted onwards from 1929 is attributable to an interaction of the relative participation rates of boys and girls in these two areas. This suggestion can be tested more thoroughly than in Masters and Beswick's analyses by using the model (5.1) in the following ways:

(1) Check the analyses based on (5.2)–(5.5) within each sex. If similar second-order properties are observed then it is evidence that the model (5.1) holds as a description of the scores of *individuals*, and that any systematic differences between subgroups formed on the basis of gender are merely gender-*linked* effects.

(2) Investigate the gender-difference of the averages of $\{Y_{ij}(1)\}$ and $\{Y_{ij}(-1)\}$ for each of $j = 1$ and 2. If these gender-based differences are of similar sign and (better still) size for the two course areas, then it is evidence that the model-based averages of Δ_i within each sex are different. Moreover, they are not related to the verbal/quantitative contrast factors $\{v_{i2}\}$. (Such evidence was supplied to the Committee that wrote *MATHEF* (1986) but not reported there.)

(3) Note in particular the correlations between the contrasts (5.3) and (5.4). If the gender-based differences observed as mode of assessment effects are attributable to verbal/quantitative contrast factors, then these correlations should differ from zero.

The presence of the factor $\{\Delta_i\}$ is of considerable concern for its effects, not only on the reference score equating procedure of section 3.5, but also when used in conjunction with any of the other course score estimation procedures of sections 3.2 to 3.4 in which scores such as $\{\widetilde{V}_i\}$ at (3.38) are used as the scores of the particular course s for which $(\alpha_s, \beta_s) = (0, 1)$ as at [1] above (3.10). This is particularly so whenever the mean squares $\sigma^2_\Delta(j) \equiv \mathrm{var}_j(\Delta_i)$ differ considerably from the quantities σ^2_j of (2.2) because the estimators b_j are affected by the ratio

$$\frac{\mathrm{var}_j(v_i) + \sigma^2_j}{\mathrm{var}_j(v_i) + \sigma^2_\Delta(j)} = \frac{1 + 1/\Gamma_j}{1 + 1/\Gamma_\Delta(j)}\,. \tag{5.7}$$

So soon as $\Gamma_\Delta(j)$ is smaller than the general range 3 to 6 for Γ_j as at (3.21), distortion of the scale estimators b_j occurs and biases the contribution of the scores Y_{ij} from the courses j concerned. It is therefore appropriate to ensure that any sub-populations whose reference test scores are used in order to establish some form of comparability across groups which otherwise have vacuous common sub-candidatures $\mathcal{C}_{jk}$ j, $k \neq s$), have their ratios $\Gamma_\Delta(\cdot)$ (over the sub-population concerned) within the range 3 to 6. Put another way, the estimate of the mean square σ^2_Δ within a sub-population can be considerably in excess of the purported measurement error associated with the reference test, and hence indicate a significant presence of mode of assessment differences $\{\Delta_i\}$; *when this is so, it is essential to consider methods of reducing this observed mean square to the order of magnitude of the measurement error so as to comply with the fundamental assumption that (3.38) holds with measurement error only.*

References

Anastasi, A. (1958). *Differential Psychology — Individual and Group Differences in Behaviour*, Third Ed. Macmillan, New York.

Cooney. G. H. (1975). Standardization procedures involving moderator variables—some theoretical considerations. *Aust. J. Educ.* **19**, 50–63.

—— (1978). A critique of standardization by bivariate adjustment—a rejoinder. *Aust. J. Educ.* **22**, 323–325.

Daley, D. J. (1985). How should NSW HSC examination marks be reported? *Independent Education* **15** (2) , 34–38.

—— (1987). *Scaling NSW HSC Marks for School-leaver Admission, February 1987.* Report to the Canberra College of Advanced Education.

—— (1988). Ranking in a one-factor model used to describe exam. marks. *Proc. Internat. Workshop/Seminar on Statistical Inference Procedures in Ranking and Selection*, Sydney, August 1987. American Sciences Press, Ohio (to appear).

—— & Eyland, E. A. (1987). The new and old HSC: Figures, facts and fantasies. *Independent Education* **17** (3), 22–25.

—— & Seneta, E. (1986). Modelling examination marks. *Aust. J. Statist.* **28**, 143–153.

Hasofer, A. M. (1978). A critique of standardization by bivariate adjustment. *Aust. J. Educ.* **22**, 319–322.

Masters, G. N. & Beswick, D. G. (1986). The construction of tertiary entrance scores: principles and issues. Technical Report, Centre for the Study of Higher Education, University of Melbourne.

MATHEF (1986). *Making Admission to Higher Education Fairer.* Report of the Committee for the Review of Tertiary Entrance Score Calculations in the Australian Capital Territory. Australian Capital Territory Schools Authority, The Australian National University, and Canberra College of Advanced Education.

Potthoff, R. F. (1982). Some issues in test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test Equating.* Academic Press, New York, 201–242.

Seneta, E. (1987). Report on the scaling of the 1986 New South Wales Higher School Certificate, University of Sydney.

Reset version of 1986 typescript

Appendix 2

Different Sex Differences from Different Modes of Assessment: Common Experiences in Three Countries

D.J. Daley

Statistics Research Section
School of Mathematical Sciences
The Australian National University

Summary

The paper reviews literature from USA, UK and Australia, all consistent with assessment based on multiple-choice methods yielding higher average male than female scores compared with teacher-assessed or external examination scores. The effect is such as to change the male : female ratio in the top half of the scores from 50 : 50 up to 60 : 40 or even more excessive.

§1. **Introduction**

The object of this paper is to summarize, with whatever quantitative detail can be given, a range of literature relating to a mode-of-assessment gender difference. What is meant by this is the following: suppose that a mixed-sex group of students finishing secondary school (age say 16 to 19) is assessed as to specific academic ability or achievement by two measures, one a multiple choice test and the other an examination or equivalent assessment of work requiring synthesis like an essay or problem-solution or homework assignment of similar format. Then the resulting assessments, when expressed on a scale with unit standard deviation, if averaged for each sex within each style of assessment, yield as gender differences quantities of which the male minus female gender difference on the multiple choice assessment is higher than the other-assessment gender difference by about 0.2 to 0.5 units.

It is quite feasible for the correlation of the two assessments to be quite high (e.g. 0.8 to 0.9), and yet for such a bias between the two assessments to exist and to be of the indicated size, having only a negligible effect on the correlation. This comment, that sizable bias and high validity may coexist, is known. (Suppose that the correlation, in the absence of bias, equals r. Then, on introducing a bias of size $2b$ (so b equals 0.1 to 0.25 in the example above), the correlation is reduced to $(r - b^2)/(1 + b^2) \approx r - (1 + r)b^2$ for small b, so for $r = 0.7$ and b in the indicated range, r is reduced from 0.7 to between 0.683 and 0.6).

§2. **Some experience in USA**

§2.1.

Breland & Griswold (1982) used data based on about 10,000 students entering Californian State Universities and Colleges in Fall 1977 and administered (i) that system's various English Placement Tests (EPT's), having already sat (ii) the Scholastic Aptitude Test (SAT) and Test of Standard Written English (TSWE). All of these except for the EPT Essay Test are multiple choice tests. The Essay Test is graded on a six-point scale by two readers, resulting in a score on a scale from 2 to 12 (with provision for a score from a third reader if the first two disagree by two or more points). Correlations between EPT-Essay scores and each of the scores listed below were in the range 0.43 to 0.48 (males) and 0.49 to 0.51 (females). Gender differences of the scores in standardized units are as shown in Table 2.1.

Breland & Griswold reported that the effect of a bias of EPT-Essay relative to any of the other tests, persists within each of four sub-ranges of the scores for each of SAT-Verbal and TSWE. They commented that

> "reviews (Breland, 1979; Linn, 1973) have suggested that, generally, women are underestimated by traditional academic tests. In other words, these studies conclude that women perform better in college . . . than traditional tests would predict."

Table 2.1

Gender differences of EPT-Essay Test and some other tests.

Test	Gender Difference		Difference of
	Test	EPT-Essay Test	Gender Diffs.
EPT-Reading	0.01	–0.36	0.37
EPT-Sentence Construction	–0.05	–0.36	0.32
EPT-Logic	0.04	–0.36	0.40
SAT-Verbal	0.10	–0.36	0.46
TSWE	–0.09	–0.36	0.27

Source: Adapted from Total, Men and Women entries in Table 3 of Breland & Griswold (1982).

Linn (1973), referring to an April 1972 conference paper of C.L. Thomas, refers to [undergraduate] college grade point averages (GPA's) for men and women in relation to their mean SAT's, noting that the regression equations predicting GPA's from SAT's in all 10 colleges underpredict the women's performance. Linn quotes a comment from the College Board Commission on Tests as being seemingly in approval of the desirability of the tests that they

> "tend to reduce the advantage that girls enjoy in grade school work, since males and females have roughly the same mean scores on the SAT."

What is really at issue here is the implied "correctness" of SAT as measuring Scholastic Aptitude as opposed to the grade school or college assessments being "correct".

It is not easy to infer from Linn's paper what is the size of the bias between the two measures. Using GPA standard deviation of 0.63 as representative of the 10 colleges, assuming a correlation of 0.6 between GPA and mean SAT, the median difference (of the 10 colleges) between predicted male and female GPA which Linn gives as 0.36 becomes about 0.34 (= 0.36(0.6/0.63)) but note the use of guesstimates!!

Breland (1979), like Anastasi (1976) but unlike Anastasi (1958), is concerned almost entirely with an extensive review of literature that gives correlations between test scores like SAT and both high school and college GPA's, especially with regard to ethnic sub-groups of the population. There is little work on the difference of gender differences we note here, except for references to Linn (1973) and an American College Test (ACT) Program study c.1973 of data from 19 colleges where

> "ACT scores and high school grades were used to predict first-semester college GPA's . . . The performance of women in the first semester of college was almost always under-predicted."

It is worth noting that Breland's (1979) review was published by the College Entrance Examination Board which two years earlier wrote (in a document attracting some public prominence), on the topic of sex differences (Wirtz, 1977, p.16):

> "Women and men have traditionally averaged about the same scores on the Verbal portion of the SAT, but there has been a marked difference in the Mathematical Averages. In 1960 the Mathematical means . . . were 465 for women, 500 for men. Twelve years later, the average for women was virtually unchanged but the average for men had dropped by 14 points (to 506). The 1977 Mathematics figures are 445 for women, 497 for men. Women represented 42.7% of the SAT-taking group in 1960 and 47.5% in 1970.
>
> The suggestion is sometimes made that the SAT is culturally biased. . . . These same differences show up in most other standardized tests . . . The test design procedures followed by ETS [ensure] that special efforts have been made to avoid the suggested prejudices. Cultural bias would appear to be more likely to affect the Verbal part than it would the Mathematical part of the test; but the differences between the averages for various ethnic groups are larger for Mathematical scores than they are for verbal scores. Although the available information is incomplete, the predictive validity of the SAT appears to be substantially the same for students in different ethnic groups and for women and men.
>
> The significant 'biases' involved here clearly go much deeper and concern the society much more than the tests . . . "

Indeed, assuming that "predictive validity" here simply refers to correlation, such validity is hardly affected by bias (see §1 above).

> "That women score lower than men on the Mathematical sections of the SAT almost unquestionably reflects more than anything else the traditional sex stereotyping of career opportunities and expectations."

Presumably, Breland, who prepared various research papers as background for Wirtz (1977), in preparing the 1979 review, was filling some of the void noted above by the CEEB. It is worth noting the caution exercised by the CEEB (p.17):

> "Realizing that even recognition of . . . group differentials risks irresponsible headlines, . . . we note the figures . . . Women's larger participation could be identified with [about 4 to 5 points] of the drop in the Mathematical average, but with none of the decline in Verbal scores."

The CEEB's reticence to associate changes in SAT average scores with changing patterns of composition of the population taking the test, contrasts with stronger emphasis on changing retention rates affecting sex differences in Australian SAT scores (Adams, 1984).

§2.2.

Stockard & Wood (1984) used scores from the California Test of Mental Maturity (which

> "as with most intelligence tests that differentiate between the sexes were omitted so that the norms show substantially equal total scores for females and males),"

and yearly grade averages (7th to 12th grades) and cumulative GPA (9th to 12th grade), excluding maths. grades beyond the 10th grade. Mean GPA's were calculated in quartiles and deciles of the

7th grade CTMM scores. Assuming equal CTMM average male and female scores at 7th grade, the mean GPA for females exceeded that for males by about $0.20/0.52 = 0.38$ standardized units. The effect persisted over the range of 7th grade CTMM scores, and within both 'working class' and 'middle class' students (a sample size of 371 students' scores who attended the school involved throughout). References to some previous work is summarized as

> "because the sexes generally score equally well on standardized achievement and intelligence tests, boys are defined as being underachievers in school more often than girls"

(because females receive higher grades than males throughout the grade school, high school and college). Certainly, such a view would appear to be folkloric amongst educational psychologists in the early '70's, for the statement:

> "It is known that the average level of performance in college is higher for females tnan for males who score the same on the same SAT exam,"

is to be found in an undergraduate text on Educational Psychology (Liebert & Poulos, 1973. p.344). This last statement could well reflect the writing of Anastasi (1958) who wrote (e.g., p.494)

> "girls generally obtain better grades than boys, even though the latter are a more select group and make a better showing on achievement tests."

Such folklore is disputed in Jensen (1980) who attempted to refute it on the basis of the sex difference in the types of courses taken by men and women. Jensen wrote (p.629) of his "hunch", and referring to Linn's (1973) work,

> "the fact is, however, that this apparent sex bias of the SAT is most probably entirely illusory, a mere artifact of the failure to control for differences in the difficulty levels and grading standards of the many college courses that enroll markedly disproportionate numbers of men and women."

He quoted a study of Hewitt & Goldman (1975) as concluding that

> "much if not most of the apparent 'overachievement' of college women is accounted for by sex differences in major field choice."

This appears to be the major basis for his rejecting the notion of sex bias in tests relative to college or high school grades. If so, then his conclusion is not supported by close examination of the overwhelming majority of data in the cited paper. Moreover, the resultant over-achievement calculated by Hewitt & Goldman is significant at the 5% level for one institution and 1% level for the other three; the data can be interpreted as yielding differences of gender differences, in standardized units as used in the present paper, of approximately 0.36, 0.11, 0.24 and 0.25.

§2.3.

Pallas & Alexander (1983), while largely concerned with relating observed SAT-M scores to 9th grade SCAT-Q scores and a plethora of other variables representing high school coursework and background, include data giving sex differences on SAT-M of 0.36 points and MATHGPA of -0.14 points (and the assumption of unit standard deviation in such scores), leading to an overall difference in gender differences of 0.50. It could be argued that MATHGPA represents an averaged assessment over grades 9–12, and that a more appropriate comparison is thus with an average of SAT-M (at 12th grade) and SCAT-Q (at 9th grade), in which case 0.36 is replaced by 0.18, and the overall difference in the assessed gender differences is 0.32. (See SUBMITTED for further detail.) The figure of 0.32 is similar to the figure of 0.38 in Stockard & Wood as above.

§2.4.

For all that it has a large annotated bibliography, and just for the decade or so starting c.1964, Maccoby & Jacklin's (1974) treatise is not much help in this present exercise. Their Table 4.1 gives just two entries that are pertinent.

Monday *et al.* (1967) gave students the American College Test (ACT), and high school grades for most of them were obtained. On the ACT, women had higher English scores, men had higher mathematics, natural science, and total composite scores. Women's high school grades were higher than those of men.

Wyer (1967) used scores on ACT Service Entrance Exam. and their first-term freshman GPA's. No sex differences were noted.

Maccoby & Jacklin wrote (their p.135):

> "It is well known that girls get better grades throughout their school years (see Maccoby (1966) including its annotated bibliography) . . . We have seen that girls do not obtain high aptitude or achievement test scores, taking all the subject-matter areas together. Hence their better grades must reflect some combination of greater effort, greater interest, and better work habits."

Sherman (1978), in her Chapter 3 in particular, notes that

> "the most recent major review of the material, Maccoby & Jacklin (1974), is a work seriously marred by conceptual, interpretive, and empirical errors"

and details both critical reviews and items with which she disagrees. Unfortunately, Sherman has nothing further to add to the obvious question as to why cognitive functioning of the two sexes may differ in a range of content areas with respect to style of assessment.

§2.5.

Veldman (1968) used a study of performance at the University of Texas and aptitude test performance and found a sex-difference of 0.23 standardized units as used here. He concluded that

> "(2) females achieved significantly higher grades relative to their aptitude test performance; (3) self-reported attitudes to work made a substantial contribution to the prediction of grades even when aptitudes were held constant; and (4) some of the sex differences in relative achievement could be shown to overlap with attitudes to work."

Caldwell & Hartnett (1967) wrote that

> "the GPA in schools and colleges has long been considered advantageous to females. Studies going back as far as 1929 reveal higher performance on school marks from elementary school through college (Anastasi, 1958, pp.492–6) . . . This paper deals with a comparison of male and female grades obtained from instructors in a variety of courses which have a built-in control for course achievement. This control is a common final examination constructed by a skilled test specialist, tailored to the course syllabus, approved by the department head, and generally demonstrated to possess appropriate psychometric characteristics (e.g. high reliability, large number of discriminating items, etc.)."

The scores were reported on a 15 point scale (of which the standard deviation is not given; assume below that it is in the range 2 to 3), and in four terms and in six subjects the differences of gender differences (F − M)(Instructor − Test) are as listed in Table 2.2.

The average figure corresponds to 0.20 to 0.30 standardized units (in terms of estimated range of the standard deviation). The authors noted

> "the consistent advantage females have in Biology and Physical Science—two non-verbal courses where essays typically play a minor part. On the other hand, in English—where instructors' ratings might be expected to favor females because of typically superior penmanship, grammar etc.—no such advantage is present."

Table 2.2

Gender difference of instructor v. test grade difference.

	Term				
Subject	1	2	3	4	Average
English	0.70	–0.02	0.25	–0.41	0.13
Behavioural Sci.	0.30	0.07	–0.16	–0.05	0.04
Biology	1.19	0.67	0.50	0.72	0.77
Physical Sci.	2.34	1.94	1.41	0.55	1.56
Mathematics	0.13	–0.13	0.77	–0.29	0.12
American Idea	0.15	0.89	1.14	1.60	0.95
Average	0.80	0.57	0.65	0.35	0.59

Source: Caldwell & Hartnett (1967).

§3. **Some experience in UK**

§3.1.

Murphy (1980) illustrates a comment in a working paper that the type of assessment techniques used within individual examinations may lead to differential gender differences, by referring to ‘O’ level Geography exam. results of the Associated Examining Board (AEB). He first notes that the differences in percentage of male and female candidates receiving A, B or C grades were in the range 0 to 2% for the years 1970–76 inclusive, and 9 to 11% for 1977–79, the two different periods corresponding to old syllabus and exam. (“written paper”), and new syllabus and exam. (both objective test questions and old-style “written paper” questions). The differences in standardized units of the male and female performances on the two parts of the paper in 1977–79 are as in Table 3.1.

Table 3.1

Gender differences on parts of ‘O’ level AEB Exam.

Year	Gender Differences		Difference of
	Objective test mark	Written paper	gender differences
1977	0.51	0.05	0.46
1978	0.53	0.07	0.46
1979	0.44	0.08	0.36

Source: By construction from data in Murphy (1980).

Murphy concluded:

> “The possible advantages gained by male candidates when written examination papers are replaced by objective tests have been discussed elsewhere, by Murphy (1978) and Dwyer (1979). In addition, this relative improvement in male performance has been noticed in a number of GCE examinations, but there appears to be no obvious reason why male candidates should do better on this form of assessment than on other forms (Murphy, 1978). The whole activity of studying sex differences in cognitive functioning has not resulted in much in the way of clear-cut findings (Maccoby & Jacklin, 1975; Peterson & Wittig, 1979) . . . One possible explanation for this particular sex difference manifestation in academic performance is the lower emphasis on verbal ability in objective test papers, as compared with more conventional written papers . . . Examinations data contain a wealth of information about the effect of sex differentiation within examination papers themselves.”

§3.2.

Hoste (1982) discusses an analysis of a biology examination for the Certificate of Secondary Education (CSE), with a theory paper comprising some multiple choice questions, some requiring sentence completion, making statements about labelled parts of diagrams, and short prose answers, and a practical paper. Relative to ability on the paper as a whole, 26 of the 133 items on the whole examination showed significant differences on performance between the sexes, 10 favouring boys (of which 9 were in the practical and multiple choice sections of the exam.), and 16 favouring girls (of which 15 were in the structured question section of the paper). There is not enough detail to calculate a realistic 'multiple choice' question advantage to boys.

Hoste refers to other work as follows:

> "Harding (1979) found that boys did better than girls on the multiple-choice section of the Nuffield 'O' level Chemistry and Physics papers. Girls on the other hand scored better on a 'conventional' Biology paper containing essay questions . . . [Carter (1952)] found that teachers' rating of girls' performance in algebra was higher than that of boys of equivalent ability as measured by an attainment test. (Although it must be noted that an alternative explanation is as valid: that the teachers' ratings were accurate and the reference test, which was probably in multiple-choice format, over-estimated boys' attainments.) . . . [re the CSE Biology paper] it may be that the more verbal format of the required answers gave them [the girls] the advantage in these questions."

From the tone of his writing, I sense that Hoste is more guarded in the strength of his conclusion than Murphy, though Hoste could have been influenced in this regard by the need for sex differences at the item level needing to be larger than the global measures used by Murphy on the Geography paper. Hoste concluded:

> "Males showed better performance on some multiple-choice items, and females on some items which required a verbal response. But these generalizations were often overridden when the subject-matter had greater relevance to the sex concerned [i.e., a question–context effect]."

§3.3.

Wood (1978) reported that the gender difference in pass rates in the University of London School Examination 'O' level English dropped from 22.2% in 1971 to 14.4% in 1972, coinciding with the introduction in that year of a multiple-choice comprehension paper. [The difference of 8% is about the same as the 9% change noted by Murphy in the Geography papers.] Wood wrote:

> "Is this just a coincidence? The answer would seem to be 'no'; our experience at the London board is that when multiple choice is introduced into an examination the boys' pass rate nearly always improves noticeably relative to the girls' pass rate, which may even decline. It does seem that multiple choice both favours boys and disadvantages girls, although, of course, it could be argued that essay tests have exactly the opposite effect."

In an earlier paper, Wood (1976) discussed differential gender responses to multiple-choice questions and "free response" questions in mathematics. A rough calculation suggests a differential gender difference at least about 0.2 units.

§3.4.

In a feasibility study of using objective testing at 16+ in Geography (cf. §3.1 above), Wiegand (1982) used a sample of 428 students from six schools.

> "The boys scored more highly on the test, confirming the findings of other objective tests [reference includes Murphy's work], although this difference was not seen to be significant."

This conclusion is ambiguous, to say the least.

§4. Some Australian experience

Daley (1985) compares gender differences on the Australian SAT scores, and continuous assessment scores aggregated into a Tertiary Entrance (TE) score. The differential gender difference for the five years 1981-85 were 0.36, 0.53, 0.17, 0.23, 0.35 (the figures for 1984 and 1985 come from Daley (personal communication)).

Each year, a new ASAT paper is devised, whereas the TE score and the components that make it up are produced by at best a slowly changing procedure (as curriculum evolves and teachers change). Thus, it is fair to conclude (as noted in Daley, 1985, and based on more than just one piece of evidence) that the gender difference of ASAT scores changes with each paper — hardly a surprising conclusion!!

Unpublished work of Daley that compares ASAT Verbal scores with English course scores, and ASAT Quantitative scores with Mathematics scores, suggests comparable gender differences in both these areas (personal communication).

§5. Quantifying the effects of "bias"

The object of this section is to illustrate numerically the effect of having a bias in one measure relative to another. At the individual level, we denote scores on two assessments, both of which have zero mean and unit standard deviation over the whole population, by X_i and Y_i for individual i, so that the true score model incorporating sex bias can be written

$$Y_i - (\text{error in } Y_i) = X_i - (\text{error in } X_i) \pm b,$$

where one sign is taken for male, the other for female. Denoting the sample male and female averages by $\overline{X}_M, \ldots, \overline{Y}_F$, and assuming the sample error means to be negligibly small, it follows that

$$\overline{Y}_M - \overline{Y}_F = \overline{X}_M - \overline{X}_F + 2b.$$

In scores such as the SAT or those occurring in Murphy (1980) or Daley (1985), it is sufficient for numerical work to assume that scores are normally distributed. Suppose that selection into another educational course is based on an academic order of merit established by ranking the scores $\{X_i\}$ or $\{Y_i\}$: when the cut-off point is the top 5% or 10% or ... or when the 'pass-rate' coincides with the top 50% of this order, what is the effect of the bias term b on the sex-ratio in this selected group? In order to be definite, assume a 50 : 50 male : female ratio throughout a mixed-sex population on the basis of the scores $\{X_i\}$. Then the entries in Table 5.1 show how the ratio changes when the bias factor b is as listed, depending on the cut-off point. Observe that the higher up the order-of-merit list is the cut-off, so the more pronounced is the effect of the bias on the proportion of the dominant sex in the selected group.

To illustrate the use of Table 5.1, refer to Murphy's example in section 3.1 above where a bias of about 0.21 was observed over three years. Assuming the scores on the two parts of the paper are weighted evenly, implies a bias overall of about 0.105 for the years 1977–79 compared to 1971–76, so that if the pass level is at the 50% cut-off, and the sex-ratio in the pass group in 1971–76 was 50 : 50, then a shift to about 54.2 : 45.8 can be expected, i.e., a change of about 8.4%, which is close to the observed figure of about 9%.

Another way of representing the effect is to use the recognition of the incorporation in X_i or Y_i of a measurement error term, and to calculate for an individual with true score at a cut-off level the probability of inclusion in the selected group when the score used is subject to bias. To a first order of approximation, such probabilities depend on the bias b and the size of the measurement error, which size we indicate by the reliability coefficient, but do not depend on the cut-off level. Obviously, the more reliable is the measure, so the more pronounced is the effect of any bias on an individual with score on the borderline.

Table 5.1

Percentage Proportion of dominant sex above selection cut-off Ievels.

Bias b	0	.05	.10	.15	.20	.25	.30
Cut-off							
5%	50	55.1	60.2	65.0	69.5	73.5	77.5
10%	50	54.4	58.7	62.9	66.9	70.6	74.1
20%	50	53.5	57.0	60.3	63.6	66.8	69.9
30%	50	52.9	55.8	58.6	61.4	64.1	66.7
40%	50	52.4	54.8	57.2	59.5	61.8	64.1
50%	50	52.0	54.0	56.0	57.9	59.9	61.8

Source: From formulae given in the Appendix.

Table 5.2

Probability of borderline individual with biased score meeting a cut-off level.

Bias b	0	.05	.10	.15	.20	.25	.30
Reliability							
0.95	.5	.589	.672	.749	.814	.868	.910
0.925	.5	.572	.642	.708	.767	.819	.863
0.90	.5	.563	.624	.682	.737	.785	.829
0.85	.5	.551	.602	.651	.697	.741	.781
0.80	.5	.544	.588	.631	.673	.712	.749
0.70	.5	.536	.572	.608	.642	.676	.708

Source: From formulae given in the Appendix.

§6. **Discussion**

In a review of this nature, it is inevitable that any literature search may be incomplete, especially through the non-appearance of reports of research which has found no significant difference of the gender differences concerned. The ideal type of study is one that records these gender differences irrespective of the size, and in this regard data such as in Murphy (1980) and Daley (1985) are valuable in providing some indication as to the variability of the bias factor which this paper has illustrated.

It is equally significant that of all the reports found in our literature search (which, while it has no claims to being comprehensive, does claim not to have suppressed any pertinent references), all except one (viz, Wyer (1967)) reported a bias favouring males under multiple choice assessment, while Wyer recorded no significant difference. As an indication of sample sizes needed, groups of at least 200 males and 200 females are needed in order to be moderately sure of finding a bias of size $b = 0.1$.

One of the limitations of gauging the extent of sex-related bias between the two measures, is to know whether it extends over the whole range (with the possible exception of floor and ceiling effects). For example, the data underlying Daley (1985) yield the graphs as shown in Figure 1 of the average male and female TE scores (Y_i in §5) relative to a given ASAT score (X_i in §5). It is clear from this diagram that regression to the mean of the average $\text{ave}_i(Y_i \mid X_i = x)$ scores occurs,

Figure 1. Illustrating sex-bias between multiple-choice ASAT scores and course-assessed (TE) scores: Mean TE scores for Females (– – –) and Males (——) within bands of ASAT scores (bandwidth = 5). 'Perfect agreement' line (TE = 3.6(ASAT) + 10) (···) also shown.

consistent with correlation between the two measures of about 0.65, that a bias between the two measures with regard to sex occurs over almost the entire range, and that there is a ceiling effect associated with high ASAT scores (where there are relatively few data points on which the plotted averages are based).

A possible statistical explanation for an apparent observed bias as in Figure 1 may be a disparate proportion between the sexes of the overall small number of individuals whose standardized scores have a difference $X_i - Y_i$ considerably in excess of twice the standard deviation of the error term (and here, 'error' refers both to measurement and model errors). The inclusion of such 'outlying' individuals in a comparison of the populations would tend to invalidate any evaluation of the presence or otherwise of an overall bias. One such data set similar to that in Daley (1985) was examined, and while the effect of any outliers was small, there appeared the suggestion that amongst such potential outliers, the females but not the males tended to have somewhat larger deviations between the verbal and quantitative sub-scale scores of the SAT concerned.

As Kingdon *et al.* (1983) imply, and as in the circumstances that lead to the work underlying Daley (1985), the existence of bias as illustrated in this review implies that any "score equating" method that relies on a multiple-choice aptitude test to place groups of students' scores from coursework or conventional examinations on a common scale, introduces a sex bias into the latter scores so soon as the gender-mix of the course-work or examination groups varies appreciably.

Regard an educational measurement as embracing behavioural responses, both to the educative process itself and to the assessment instrument(s) being used. Then it is quite plausible a priori that there may be interactions with those responses attributable to any different behavioural patterns of the sexes, in spite of any attempts in those assessments to focus purely on matters reflecting mental processes. If mental processing affects the emotional disposition of an individual, and if emotional dispositions vary between the sexes, then a sex-related interaction effect on mental processing can be anticipated.

Indeed, in terms of possible differences in cognitive functioning of the two sexes, the assessment of individuals on the same substantive material through a variety of instruments may point to differences in such functioning. Papers such as those of Wood (1976, 1978) and Hoste (1982) may be useful in this regard through their discussion cf individual items.

Appendix: Formulae for the Tables of Section 5.

Let $\Psi(x)$ denote the tail area of a standard normal distribution, meaning, that if X is normally distributed with zero mean and unit standard deviation, then $\Pr\{X > x\} = \Psi(x)$. Table 5.1 tabulates $\Psi(x-b)/[\Psi(x-b)+\Psi(x+b)]$ for the values x satisfying $\Psi(x) = 0.05, 0.1, 0.2, \ldots, 0.5$.

For Table 5.2, recall that with the true score model for which

$$\text{observed score} = \text{true score} + \text{error},$$

the reliability R is given by

$$R = \text{var(true score)}\big/\,\text{var(observed score)}.$$

The error standard deviation s is thus equal to $\sqrt{1-R}$. Assuming a normal distribution for the "error" variable, Table 5.2 tabulates

$$\Psi\big(-b/\sqrt{1-R}\big).$$

Obviously the numerical results arising from these formulae should not be interpreted literally, but rather as indicative of changes that occur depending on the model parameter values.

References

Adams, R.J. (1984). *Sex Bias in ASAT?* ACER Research Monograph No. 24. Hawthorn, Vic., Australia: Australian Council for Educational Research.

Anastasi, A. (1958). *Differential Psychology — Individual and Group Differences in Behaviour*, Third Edition. New York : Macmillan.

Anastasi, A. (1976). *Psychological Testing*, Fourth Edition. New York: Macmillan.

Breland, H.M. (1979). *Population validity and college entrance measures.* College Board Research Monograph No. 8. New York: College Entrance Examination Board.

Breland, H.M. & Griswold, P.A. (1982). Use of a performance test as a criterion in a differential validity study. *J. Educ. Psych.* **74**, 713–721.

Caldwell, E. & Hartnett, R. (1967). Sex bias in college grading? *J. Educ. Meas.* **4**, 129–132.

Carter, R.S. (1952). How invalid are marks assigned by teachers? *J. Educ. Psych.* **43**, 218–228.

Daley, D.J. (1985). Standardization by bivariate adjustment of internal assessments: sex bias and other statistical matters. *Australian J. Educ.* **27**, 231–247.

Dwyer, C.A. (1979). The role of tests and their construction in producing sex-related differences. pp.335–353 in Wittig & Peterson (1979).

Harding, J. (1979). Sex differences in examination performance at 16+. *Physics Educ.* **14**, 280–284.

Hewitt, B.N. & Goldman, R.D. (1975). Occam's razor slices through the myth that college women overachieve. *J. Educ. Psych.* **67**, 323–330.

Hoste, R. (1982). Sex differences and similarities in performance in a CSE biology examination. *Educ. Studies* **8**, 141–153.

Jensen, A.R. (1980). *Bias in Mental Testing*. London: Methuen.

Kingdon, J.M., French, S., Pierce, G.E. & Woodthorpe, A.J. (1983). Awarding grades on differentiated papers in school examinations at 16 plus. *Educ. Res.* **25**, 220–229.

Liebert, R.M. & Poulos, R.W. (Eds.) (1973). *Educational Psychology — A Contemporary View.* Del Mar, Calif.: Communications Research Machines Inc. Books.

Linn, R.L. (1973). Fair test use in selection. *Review Educ. Res.* **43**, 139–161.

Maccoby, E.E. & Jacklin, C.N. (1974). *The Psychology of Sex Differences*. Stanford: Stanford Univ. P., and (1975) London: Oxford Univ. P.

Monday, L.A., Hout, D.P., & Lutz, S.W. (1967). *College Student Profiles: American College Testing Program.* Iowa City: ACT Publications.

Murphy, R.J.L. (1978). Sex differences in objective test performance. Unpublished AEB Research Report, RAC/56.

Murphy, R.J.L. (1980). Sex differences in GCE examination entry statistics and success rates. *Educ. Studies* **6**, 169–178.

Pallas, A.M. & Alexander, K.L. (1983). Sex differences in quantitative SAT performance: new evidence on the differential coursework hypothesis. *Amer. Educ. Res. J.* **20**, 165–182.

Peterson, A.C. & Wittig, M.A. (1979). Sex-related differences in cognitive functioning: an overview. pp.1–17 in Wittig & Peterson (1979).

Sherman, J.A. (1978). *Sex-related Cognitive Differences: An Essay on Theory and Evidence.* Springfield, Ill.: Charles C. Thomas.

Stockard, J. & Wood, J.W. (1984). The myth of female underachievement: a re-examination of sex differences in academic underachievement. *Amer. Educ. Res. J.* **21**, 825–838.

Veldman, D.J. (1968). Effects of sex, aptitudes and attitudes on the academic achievement of college freshmen. *J. Educ. Meas.* **5**, 245–249.

Wiegand, P. (1982). Objective testing in Geography at 16+. *Geography* **67**, 332–336.

Wirtz, W. (1977). *On Further Examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline.* New York: College Entrance Examination Board.

Wittig, M.A. & Peterson, A.C. (eds.) (1979). *Sex-Related Differences in Cognitive Functioning. Developmental Issues.* New York: Academic Press.

Wood, R. (1976). Sex differences in mathematics attainment at GCE ordinary level. *Educ. Studies* **2**, 141–160.

Wood, R. (1978). Sex differences in answers to Enqlish language comprehension items. *Educ. Studies* **4**, 157–165.

Wyer, R.S. Jr. (1967). Behavioural correlates of academic achievement: conformity under achievement- and affiliation-incentive conditions. *J. Personality and Social Psych.* **6**, 255–263.

Replicate of Figure 1 based on ACT 1997, 1998, 2000 and 2001 datasets (mixed-sex colleges). Illustrating sex-bias between multiple-choice ASAT scores and course-assessed (TE) scores: Mean TE scores for Females (– – –) and Males (——) within bands of ASAT scores (bandwidth = 5). 'Perfect agreement' line (TE = ASAT + 2) (· · ·).

Complementary version of Figure 1 based on ACT 1997, 1998, 2000 and 2001 datasets (mixed-sex colleges). Illustrating sex-bias between multiple-choice ASAT scores and course-assessed (TE) scores: Mean ASAT scores for Females (– – –) and Males (——) within bands of TE scores (bandwidth = 10 at top, = 5 elsewhere). ‘Perfect agreement’ line (TE = ASAT + 2) (· · ·).