APPLIED MEASUREMENT IN EDUCATION, 9(4), 323–353


# Analysis of Profiles of Students Applying for Entrance to Universities

Jim Tognolini
*Educational Testing Centre*
*University of New South Wales*

David Andrich
*School of Education*
*Murdoch University*

Applicants to universities present profiles of performance in a variety of relevant content areas as evidence for selection. Even though profiles of different applicants may involve different content areas, the applicants may be competing for places to the same university and even to the same program of study. In that case, and if there are more eligible applicants than there are places, universities must reconcile these different profiles in order to make comparisons among them. When there is a small number of applicants, then these comparisons may be carried out qualitatively; when the number of applicants runs into the order of thousands and there is a short time in which to make the offers, some quantitative analysis is required. This quantitative analysis usually involves aggregating the components of each profile in order to form a single score from which comparisons among applicants can be made readily. Legitimate concerns can be raised regarding forming simple aggregates from diverse components of profiles, but despite these concerns, the practical problem of making relatively rapid decisions means that these concerns are generally not addressed. The premise of this article is that profiles will be more or less consistent among the components and that although some profiles may not be, a great number of others may be, summarized adequately by a single score. It is shown that by applying the principles of latent trait test theory at the level of tests, it is possible to rank order a

Requests for reprints should be sent to Jim Tognolini, Educational Testing Centre, University of New South Wales, Kensington, New South Wales 2033, Australia.

set of profiles in terms of the adequacy with which they are summarized by a single score, and that as a result, only a subset of the original profiles may require a qualitative analysis. The application of the procedure is illustrated with a random sample of 577 profiles from a population of 12,314, which were presented for selection into universities in Western Australia in 1986.

Increasing school populations in the postcompulsory years of schooling in Australia together with a restricted number of places in tertiary institutions has heightened awareness about the processes involved in the selection and rejection of applicants. In this article, we restrict ourselves to tertiary institutions that are universities. Applicants to universities present profiles of performance in a variety of relevant content areas as evidence for selection. Even though profiles of different applicants may involve different content areas, the applicants may be competing for places to the same institution and even to the same program of study. Inevitably, when there are more applicants than there are positions, these profiles need to be compared. If, as in the applications for some positions in employment, there are relatively few applicants, then a qualitative analysis and comparisons among the profiles can be carried out. However, when the number of applicants runs into thousands, and decisions need to be made within a few weeks, this kind of study of profiles is not possible, and some kind of quantitative analysis is applied. Such a quantitative analysis usually involves forming an aggregate or a single score from the components of each profile in a more or less sophisticated way. Concerns are often expressed that such a single score cannot capture the richness of the information in each profile (Matters, 1987; Viviani, 1990).

Our hypothesis in this article is that different profiles can be summarized by a single score more or less precisely and that although some profiles cannot be, many others can be summarized by a single score. In this article, we use a measurement model of modern latent trait test theory at the level of tests in content areas in order to provide two rank orders of the profiles, the first according to a single score that summarizes a level of performance and the second, the degree to which this single score captures the total information in each profile. The second rank ordering identifies those profiles that require closer qualitative analysis. Despite the limits on resources for considering each profile qualitatively, it might be possible to target some resources to carry out a qualitative analysis of those identified by the suggested procedure.

Although, in this article, we use the case of entrance into public universities in the state of Western Australia, which itself is illustrative of selection procedures in other states of Australia, the rationale and methodology is not restricted to such situations. The procedure can, in principle, be used to analyze and order profiles for ranking and selecting in a greater variety of similar situations. Nevertheless, the sheer volume of candidates and the time that is available to carry out the selections means that case of entrance to universities is an important end in itself, and accordingly, in this article we continue to refer to this case.

## PROFILE AND AGGREGATION OF THEIR COMPONENTS

Before proceeding to the model and its application, three further preliminary issues are considered briefly: first, the essential aspects of the procedures for selection into universities in Western Australia, which helps place the theme of the article in context; second, the issue of unidimensionality that is raised by the rank ordering of profiles according to a single score; and third, validity, in which the selection and rejection of applicants implies that those selected are more likely to succeed than those rejected.

### Selection Into Universities in Western Australia

The majority of entrants into the four public universities in Western Australia follow studies of the Year 12 curriculum at secondary school, and the case refers to that class of applicants. A small minority of mature-age students enter by other various procedures, which vary from university to university. At the time the data in the example were collected, students had to study a minimum of 6 content areas in which a minimum of 3 had to belong to a tertiary entrance set of 28 designated areas. The syllabi of these content areas are approved in part by committees that have university representatives. The other content areas are influenced less, if at all, by universities. Within the tertiary entrance set, students had to include at least one area from two broadly based domains of study, one designated as the "Humanities/Social Studies" and the other as "Quantitative/Science Studies," and they had to study one of two areas of English, "English" and "English Literature." The requirement of a minimum of one content area from each of the two general domains was designed to ensure some breadth of study, whereas the possibility to study up to five content areas in the same domain permits also substantial depth of study. The only change to this selection procedure in 1995 is that the minimum number of content areas in the tertiary entrance set is four.

Every student who met the requirements of having at least three content areas from a tertiary entrance set with at least one from each domain, had a single tertiary entrance score (TES) calculated. Students were offered places according to this ranking among the applicants into each university and each program of study. The TES is calculated in the following way. First, each content area has a score with a maximum of 100 rounded to the nearest integer, which is composed equally from the sum of a school-based assessment and an external examination. The school-based component in each school is scaled linearly against the external examination to overcome possible differences in standards among schools. Second, the scores in each content area are scaled using the Australian Scholastic Aptitude Test (ASAT), which each student takes. The ASAT is a 100-item multiple-choice test covering the areas of humanities, social sciences, mathematics, and the physical sciences, but is not tied to any specific Year 12 syllabus. This scaling involves the standard procedure of equating the distribution of scores of those students in each content area with the distribution of the scores of the same students on ASAT, with the maximum scaled score constrained to 100. Thus, if the scores on ASAT of students in content area *X* are greater than those of students in content area *Y*, then irrespective of the original scores on *X* and *Y*, the scaled scores of students in *X* will, on average, be higher than the scaled scores of students in *Y*. Of course, there may be students who take both the content areas *X* and *Y*. Third, the scaled scores are summed, multiplied by five, and divided by the number of tertiary entrance content areas a student studied, and this gives a maximum possible score of 500 to each student. Finally, 10% of a student's score on ASAT is taken and added to the sum of the scores on the content areas to give a maximum score of 510. The score on each component of the profile, which can range from 0 to 100, reflects a distinct content area, and with such a range of scores, each can be treated as a test score.

## Unidimensionality

In forming an aggregate from a profile of scores, the formal assumption is that the responses are unidimensional. At first glance, profiles of scores are unlikely to be unidimensional. However, in order to anticipate the perspective of this article, we note that whenever two profiles are compared for the purpose of selecting one ahead of the other for some purpose, the profiles are aggregated in some sense, mapped onto a single relevant hypothesized continuum, and compared. This may be carried out informally as when two or more applicants are considered for a single position in employment, in which many and varied components such as responsibility, level of general academic qualification, specific studies within the general qualification, work experience, and so on, are considered relevant. Even though these are very different substantive components, they are related to the requirements of the position, and therefore they must be integrated in each profile, the profiles must be ranked, and the person with highest ranked profile must be selected. Thus, the requirement that components of a profile be integrated into single indexes is not dependent on the components being homogeneous in some sense; instead, it is dependent on the requirements of the position, which itself requires the integration of the components in order to execute the work in the position.

Returning to the case of entrance to universities in Western Australia and with the aforementioned perspective in place, we note the following points: First, the present procedure of forming a single score already implies a kind of unidimensionality. Second, in anticipation of the more detailed discussion later in the article, we stress that unidimensionality is not an absolute but that it is a relative concept and that a level of statistical precision can always be specified at which any empirical data set will not conform to a chosen unidimensional model. This is because, as is well known, no model can be a perfect summary of an empirical data set, and all data are multidimensional at some level of precision. Moreover, there is not one single test of unidimensionality that is necessary and sufficient to assert that the data are unidimensional—each test of fit checks for different violations of the model at different levels of power. On the other hand, there are data sets that can be said to conform to unidimensionality for the purpose for which the data are collected, and it is acceptable to treat the data as unidimensional for that purpose (Thurstone, 1928). Statistical indexes can be used to quantify the degree to which the data conform to a unidimensional model, but the decision that data are unidimensional for the purpose required, as will be illustrated later in the article and is suggested in the previous paragraphs, is as much a policy and resource matter as it is a statistical one. The application of statistical models reduces the data and summarizes the information for easier consideration with respect to the policy and substantive requirements, but eventually the policy and substantive requirements will prevail.

Third, there are different levels of scale at which unidimensionality may be considered. In the context just described in Western Australia, one is at the level of each school-based assessment and the external examination in each content area. Thus, the school-based assessment is a single score that is a summary of many school-based assessments in many different components of study over a year in areas of study such as history, physics, mathematics, English, and so on. Likewise, the external assessment is a single score based on more than one test question in each examination lasting 3 hr. At the next level, a single score is formed on the basis of the school-based assessment and the external examination. Although the school-based assessment and the external examination are related to the same content area, both are explicitly used because each is better suited to assessing different aspects of study in the content area. For example, sustained project or laboratory work cannot be well assessed in an external examination, whereas an external examination is not hampered by possible differences in standards. Finally, there is the level of the TES itself, which is the sum of the different component areas.

Although this final level of unidimensionality at the TES level is in some ways no more than an extension of the other two levels, it does contain some special circumstances. Technically, these special circumstances are distinguished by the ideas of equating and scaling. In the former, scores in similar content areas are equated; in the latter, scores from different content areas are made relatively equivalent in terms of some population. Clearly, the case described in this article has an element of both scaling and equating. Nevertheless, if the function of the latter is to take a further step of forming a single score, as it is in the case of entrance to universities in Western Australia, then the idea of relative unidimensionality still applies. Selection to universities in Western Australia is based explicitly on a single score even though it also requires performance in content areas with some heterogeneity analogous to the heterogeneous components for a position in employment referred to earlier. For example, and to be specific, a selection decision might require a comparison between a profile containing the content areas of History, English Literature, French, and Chemistry and a profile containing Mathematics 1, Biology, English, and Physics or vice-versa, content areas that appear in the example in Section 4.

This discussion leads to a distinction between profiles homogeneous in content and profiles homogeneous in the values of the components. The latter refers to those profiles in which, after scaling, the components have similar values, that is, if one component is relatively high, then the other components are similarly high. Such profiles can be summarized statistically by a single score even if the components are not homogeneous in content, and in order to emphasize this feature, these profiles will be termed *consistent.* Profiles that do not have similar values in their components and which, therefore, cannot be summarized by a single score, will be termed *inconsistent.* Consistency, however, is a matter of degree.

There is an important factor in the case of students applying for positions in universities that will tend to make the profiles consistent, even though not every applicant has scores on each content area and many applicants have content areas not common to those of many other applicants, and the content areas are not homogeneous. This factor is that students select a small subset of the available content areas for study, and within the rules for selection, they will, in general, choose those areas in which they are most likely to succeed. Many able students choose heterogeneous content areas because they are equally likely to succeed in either domain. Those who are less comfortable with heterogeneous content areas will, within the rules that require that they must take one content area from each of the humanities/social studies and from mathematics/science, again, tend to choose the content area in which they are most likely to succeed. Moreover, they will make a special effort, including obtaining extra support in their studies in order to overcome any relative weakness compared with the rest of their profiles, thus enhancing more consistent profiles.

The previous analysis leads to the perspective that unidimensionality is not investigated primarily on the basis that the scores in the matrix have, in general, a uniformly high intercorrelation, but on the basis that it is investigated at the level of the individual profile, which is made up of a small subset of self-selected content areas. Although the two are related in the sense that if the correlation matrix among the content areas were uniformly low, then very few profiles would be consistent, it is considered that the self-selection of a small subset of the content areas by each person and knowledge of the rules for selection into universities, which governs the preparation for study in these areas, should tend to provide relatively consistent profiles. In this article, we summarize the scores of consistent profiles and identify inconsistent profiles; it is, therefore, considered redundant to investigate the basic correlation matrix as a criterion for aggregation (Mislevy, 1985), which, in any case, has substantial missing data. In addition to be being redundant, it can even be somewhat misleading for the case made in this article. Some low correlations among the content areas could lead readily to the conclusion that scores on the content areas should not be summed for any profile—however, despite a relatively low correlation across all the persons in a sample, there may, nevertheless, be many individual profiles for which a single index is meaningful. Finally, on this point, conformity to a unidimensional model is a more specific and stronger criterion for use of a single score than the intercorrelations among the content areas. As will be seen in the example, there is one content area that shows poorer fit to a unidimensional model than the others.

Because the case for expecting consistent profiles rests, in part, on students selecting their areas of study, this could be a limitation of the application. The degree to which it is a limitation in any particular case is an empirical question. Again, as will be seen in the example, the one content area that is identified as not fitting the unidimensional model as well as the other areas is one in which there was effectively less choice.

## Validity

In the same way as unidimensionality, validity is a relative concept. Thus, an existing procedure exists, and the question of validity of a new procedure becomes one of whether the suggested procedure is more valid than the existing one. The validity of the existing procedure is checked from time to time in Australia by studying the performances of students in universities as a function of their TES

(Dunn, 1982; Everett & Robins, 1991; West, 1985). This shows a modest but positive relation, with variation among programs and types of schools that prepare students for university entry. That the relation is modest, although positive, is not remarkable. Many factors attenuate the relations, for example, the differences in methods of teaching in schools from universities, differences in social organizations, changing interests of students including loss of interest in studies, the relative homogeneity of students on the TES in many programs of study due to selection, and especially the lack of reliability of both entrance examinations and criterion examinations at universities. However, the issue here is not simply one of predictive validity—it is already considered that performance in the content areas and the learning processes required to manifest those performances are relevant more or less specifically for preparation for university study. The more significant issue, given that the studies in secondary school are generally taken as relevant preparation for study in universities, as just described, is that of justice or fairness of selection based on the stated rules for selection. It is from this perspective of fairness in order not to penalize students studying content areas in which it might be difficult to obtain high scores and to encourage them to study these content areas as much as it is from the perspective of predictive validity that the scores in the various content areas are scaled, at present, using ASAT before they are summed. Of course, these procedures might have the consequence of increasing predictive validity, and they might be justified this way in part, but that is not the sole motivation for the scaling. Suggestions to improve predictive validity by calculating different indexes for entry to different programs have been made (Masters & Beswick, 1986; Viviani, 1990), but this tends to work against the principle of delaying specialization of studies.

Although we have argued that the summing of the scores does not require homogeneity of content areas, this summing does have another technical implication—it implies that the sum of scores of content areas contains all of the meaningful information in the profile. Thus, even though predictive validity is the underlying theme in university selection, the focus in this article is on the logic and means for identifying those profiles to which administrators might give special attention because all of the meaningful information in a profile is not contained in the total score. This perspective may also be taken from the same point of fairness—a single score should not simply be used for selection when the total score does not contain all of the information in the profile. There are, of course, the questions of whether the further evidence, beyond the total score, is relevant and to what degree administrators can make use of it. This clearly depends on circumstances in particular universities and is beyond the scope of this article. The point of this article is to show how profiles may be ranked according to a single score and then ranked again on how well this single score summarizes the statistical information in the profile. This methodological article provides the opportunity for subsequent substantive studies of predictive validity to be carried out.

## LATENT TRAIT MEASUREMENT MODEL

The latent trait measurement model that is applied is an extension of the simple logistic model (SLM) of Rasch (1960/1980), studied extensively (e.g., Andrich, 1988; Fischer, 1981; Wright, 1968, 1977; Wright & Douglas, 1977) for dichotomously scored items to those that may have more than two categories. For the purpose of this article, it is termed the extended logistic model (ELM). Although the ELM is not derived in full in this article, for completeness, it is reviewed briefly and then elaborated for the analysis required in this article. More complete steps in the derivations can be traced through Rasch (1961), Andersen (1977), and Andrich (1978, 1985). The justification for the use of this model will be made following its presentation.

### Extended Logistic Model

The SLM takes the form

$$\Pr(X_{ni} = x) = \frac{1}{\gamma_{ni}} \exp\{x(\beta_n - \delta_i)\}, \tag{1}$$

where $\beta_n$ is the latent ability of person *n*, $\delta_i$ is the difficulty of item *i*, $X_{ni}$, $x \in \{0, 1\}$ is a dichotomous variable and $\gamma_{ni} = 1 + \exp(\beta_n - \delta_i)$ is a normalizing factor that ensures that $\Pr\{0\} + \Pr\{1\} = 1$. A characteristic feature of the SLM is that the total scores $r_n = \Sigma_i x_{ni}$ and $s_i = \Sigma_n x_{ni}$ are jointly sufficient for the parameters $\beta_n$ and $\delta_i$. In particular, all of the information for the parameter $\beta_n$ is contained in the total score; that is, given the total score, the probability of the pattern of responses is independent of the parameter $\beta_n$. Statistical sufficiency is the significant feature of the model for implementing the task of rank ordering the profiles according to the degree to which the total score summarizes the information in the profile. It will be elaborated upon further when the ELM is presented.

The ELM takes the form

$$\Pr\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp\{\kappa_{xi} + x(\beta_n - \delta_i)\}, \tag{2}$$

where $\beta_n$ and $\delta_i$ retain the same meaning as in the SLM, $X_{ni}$, $x \in \{0, 1, 2, \dots, m\}$ reflect graded responses in which the maximum score is *m*, and

$\gamma_{ni} = 1 + \sum_{k=1}^{m} \exp\{\kappa_{ki} + k(\beta_n - \delta_i)\}$ is a normalizing factor that ensures that $\sum_{x=0}^{m} \Pr\{x\} = 1$. The $\kappa_{xi}$, $x = 0, 1, 2, \dots, m$ are known as category coefficients and are defined (Andrich, 1978) by the recursive relations

$$\begin{aligned} \kappa_{xi} &= 0, x = 0; \\ \kappa_{xi} &= -t_{1i} - t_{2i} - t_{3i} \dots -t_{xi}; \\ \kappa_{xi} &= 0, x = m; \end{aligned} \tag{3}$$

where $\tau_{1i}, \tau_{2i}, \tau_{3i}, \dots, \tau_{xi}, \dots, \tau_{mi}$ are $m$ successive thresholds that divide the $m + 1$ graded categories on the latent continuum.

The initial work with the ELM focused on analyzing data obtained from graded responses to attitude items designed in the Likert tradition and achievement items in which the maximum score is generally of the order of 5 or 6 and no greater than a single digit number (Wright & Masters, 1982). In such situations, it is possible and relevant to estimate all of the $m$ thresholds.

In the case that the model is applied at the level of tests for content areas, in which the maximum score $m$ is of the order of 50 or 100 rather than at the level of the items in which the maximum score is a single digit number, it may be extravagant to estimate every threshold $m$. In addition, it would be technically difficult to estimate all of the thresholds if all of the possible scores did not have responses on all of the possible scores; for example, in tests for content areas with possible scores from 0 to 100, one seldom finds scores below 20 and above 95. Therefore, it may be relevant and necessary to reparameterize the thresholds to a smaller number of parameters. Such a reparameterization has been formalized by Andrich (1985) and generalized by Pedler (1987), in which the category coefficients are expressed in terms of three principal component parameters according to the relation

$$\kappa_{xi} = x(m - x)\theta_i + x(m - x)(2x - m)\eta_i + x(m - x)(5x^2 - 5xm + m^2 + 1)\psi_i\,, \tag{4}$$

where $\theta_i$, $\eta_i$, and $\psi_i$ are parameters that characterize the dispersion, skewness, and kurtosis, respectively, of the distribution of a score on any particular content area. Note that the model does not characterize the spread, skewness, and kurtosis parameter of the distribution of scores across persons on any content area $i$, but of the theoretical distribution of responses of a single person.

The reduction of the number of parameters to be estimated for each content area from 98 to 4 (including the difficulty parameter $\delta_i$) is substantial when the maximum score is 100. It is evident that the successive coefficients are polynomials of increasing order being with the second order, which for efficiency can be summarized as

$$g_2(x) = x(m - x);\ g_3(x) = x(m - x)(2x - m);\ g_4(x) = x(m - x)(5x^2 - 5xm + m^2 + 1).$$

If we also define $g_1(x) = x$ for completeness, which is a first order (linear) polynomial and therefore fits into the polynomial pattern, then the model may be written as

$$\Pr\{X_{ni} = x\} = \frac{1}{\gamma_{ni}}\exp[x\beta_n - g_1(x)\delta_i + g_2(x)\theta_i + g_3(x)\eta_i + g_4(x)\psi_i] \tag{5}$$

in which the coefficient of the person parameter $\beta_n$ is simply the score $x$.

Because of the large values of the coefficients $g_1(x)$, $g_2(x)$, $g_3(x)$, $g_4(x)$, when the maximum score is of the order of 100, the exponents in the estimation get too large for practical estimation. Therefore, the model is reconstructed as follows: The coefficient of the person parameter is divided by $m$, and the successive item coefficients are divided respectively by $m$, $m^2$, $m^3$, and $m^4$ to give coefficients $g_1^*(x) = \frac{g_1(x)}{m}, g_2^*(x) = \frac{g_2(x)}{m^2}, g_3^*(x) = \frac{g_3(x)}{m^3}, g_4^*(x) = \frac{g_4(x)}{m^4}$. This rescaling of the coefficients rescales the parameters of Equation 5 correspondingly to give the model

$$\Pr\{X_{ni} = x\} = \frac{1}{\gamma_{ni}}\exp[(x/m)\beta_n^* - g_1^*(x)\delta_i^* + g_2^*(x)\theta_i^* + g_3^*(x)\eta_i^* + g_4^*(x)\psi_i^*], \tag{6}$$

where $\beta_n^*$ is $m\beta_n$, $\delta_i^*$ is $m\delta_i$, $\theta_i^*$ is $m^2\theta$, $\eta_i^*$ is $m^3\eta$, and $\psi_i^*$ is $m^4\psi$.

The model with the rescaled parameters is not identical to the model of Equation 4 with the original parameters because the transformation of the coefficients is not linear. Nevertheless, when the maximum score $m$ is the same for each content area $i$, then all of the properties of the original model, in particular the properties of sufficiency, hold. Therefore, for simplicity, and because it has the same structure, the model of Equation 6 will still be referred to as the ELM. The degree to which the data and, in particular, the profiles fit this model is an empirical question.

## Sufficient Statistics in the ELM

The method of estimation of the parameters of the content areas is outlined in Andrich (1985) and is not entered into in this article. It is a pairwise estimation procedure that eliminates the person parameter $\beta_n^*$ of each person. Thus no distribution, such as the normal needs to be assumed. Estimates of the difficulty, dispersion, skewness, and kurtosis parameters effectively provide the scaling that is achieved by using ASAT. This will be illustrated in the example. Given the estimated values of the parameters of the content areas, the person parameter $\beta_n^*$ is solved by the direct maximum likelihood equation

$$r_n = m\sum_{i=1}^{I_n}\sum_{x=0}^{m} x\Pr\{x\} = m\sum_{i=1}^{I_n}\sum_{x=0}^{m}\frac{1}{\gamma_{ni}} x\exp[(x/m)\beta_n^* - g_1^*(x)\delta_i^*$$

$$+g_2^*(x)\theta_i^* + g_3^*(x)\eta_i^* + g_4^*(x)\psi_i^*]. \qquad (7)$$

Of major significance in the ELM is that the value of the random variable appears in the expression of the distribution and that the statistic $r_n = \sum_{i=1}^{I_n} x_{ni}$, (which appears on its own on the left of Equation 7), is sufficient for the person parameter $\beta_n^*$. This has been demonstrated thoroughly in the literature (e.g., Andersen, 1977; Andrich, 1985; Rasch, 1961; Wright & Masters, 1982) and need not be repeated in this article. This parameter characterizes the location of a person on a single hypothesized latent continuum, termed in this article as *latent ability* for convenience, and because the person is characterized by a single parameter $\beta_n^*$, the profiles will be ordered according to the estimate of this single parameter, which is directly related to the total score. This is the ordering required for university selection. The sufficiency of the statistic $r_n$ means that all of the meaningful information in the profile for an individual person is contained in the total score, and, if the data fit the model, then all of the remaining information is error. This is the key characteristic that makes the model relevant for purposes of this article and that excludes other models, such as the graded response model of Samejima (1969) and the extended partial credit model described by Muriaki (1992), which do not have sufficient statistics for the person parameter. These models are more relevant to characterizing and describing data, but no statistic exists in these models that summarizes all of the meaningful information for a single location of the parameter $\beta_n^*$. Clearly, there is substantive information in the profiles, for example, the content areas, but the universities in Western Australia (except in some exceptional cases in which some prerequisites are listed) consider only the statistical information within the rules of the tertiary entrance set of content areas. Our point in this article is to indicate which profiles should be studied more closely because there is statistical, and therefore possibly relevant, substantive information in the profile.

Sufficiency, however, is a property of the ELM and not of the data. The total score is sufficient for $\beta_n^*$ in any particular profile only if the components of the profile accord with the ELM. On the other hand, if the components of the profile do not accord with the ELM, it means that the total score is not sufficient for the estimate of the parameter $\beta_n^*$, and therefore, that there is further statistical information in the full profile. This is a subtle but central point to the application of the model in this article: To the degree that the components of a profile conform to the ELM, all of the meaningful information is contained in the total score; to the degree that they do not so conform, that total score does not contain all of the information. This implies that a statistic that characterizes the fit of each profile to the ELM can be used to order the profiles according to the degree to which each profile is summarized by the total score. Accordingly, we form a fit statistic at the level of each profile.

Before proceeding to construct this statistic, two further points that further clarify the perspective of this article are noted. First, it is possible to construct a statistic that emphasizes the accord between the data and the model at the level of the content areas, and we do that. This is analogous to using fit statistics at the level of items in the item analysis of a test. However, although this is useful to orient the profile analysis in the case considered in this article, it does not lead to any practical consequences. It does not, for example, permit the elimination of a content area in any profile if it were found that the particular content area fitted relatively poorly to the model. Furthermore, if any particular content area did fit relatively poorly to the model, then it will be reflected in the profiles that show misfit to the model—profiles that have that content area included will tend to show misfit more than those which do not have that content area included. Thus, the misfit of a content area is accounted for, but at the level of the profile. Finally, even though the content area may misfit at some statistical level, most profiles may still accord with the model. Nevertheless, in order to get an orientation to this aspect of the data, a test of fit between the data and the model at the level of content areas is provided in the example.

Second, the focus on the test of fit is the conformity of the data to the model and not the other way around. Thus, it would miss the point in this article to search for models that fit the data better but that destroyed sufficiency because such models, as noted earlier, would not have a statistic that captured all of the meaningful information in the profile. Thus, the model makes explicit the requirements of the data if the profiles are to be summarized by a single score, and because the model is an expression of these requirements, it is chosen independently of the data.

## The Fit Statistic for the Profiles

The statistic to formalize the fit of the profile to the model is relatively straightforward and has been employed in various forms and adaptations as a person fit index in the Rasch class of models (Linacre & Wright, 1994; Ludlow, 1985; Ryan, 1980; Smith, 1986, 1988; Wright & Stone, 1979), and it has been applied in substantive studies (e.g., Phillips, 1986). It is based on the standardized residual between the observed score and the expected score in the test of each content area, given the estimated parameters of the content areas and the estimated location of the profile. Specifically,

$$Z_{ni} = \frac{x_{ni} - E[X_{ni}]}{\sqrt{V[X_{ni}]}}, \tag{8}$$

where

$$E[X_{ni}] = \sum_{x=0}^{m} x_{ni} \Pr\{x_{ni}\}, V[X_{ni}] = \sum_{x=0}^{m} x_{ni}^2 \Pr\{x_{ni}\} - (\sum_{x=0}^{m} x_{ni} \Pr\{x_{ni}\})^2.$$

This standardized residual can be used to highlight those content areas of a profile that are inconsistent with the a total score containing all of the relevant information in the profile—the greater the residual, the greater the misfit to the model. To form a single index, these residuals are squared and summed across the components of a profile, analogously to forming a chi-square statistic from independent standardized normal deviates:

$$\chi_n^2 = \sum_{i=1}^{I_n} Z_{ni}^2, \tag{9}$$

where $I_n$ is the number of content areas in the profile $n$. This is the outfit statistic described in Linacre and Wright (1994) and is sensitive to extreme values in components of the profile. It is this statistic that is used to order the profiles according to their fit to the model.

This statistic invites the calculation of an index that indicates how seriously the profile violates the model. Given that the statistic of Equation 9 takes the form of a chi-square, it can be compared to the corresponding degrees of freedom and a probability that the statistic, or greater, violates the model by chance calculated. The rationale for the calculation of the degrees of freedom is shown in the Appendix. A useful index for quick reference is the ratio of the observed $\chi_n^2$ statistic to the degrees of freedom $f_n$:$\chi_n^2 / f_n$. Because, theoretically, the expected value of a chi-square statistic is its degrees of freedom, this ratio will be close to 1.0 if the data conform to the model, and it is referred to as a fit mean square statistic (Linacre & Wright, 1994).

In the case in which $m$ is large, this statistic should be even more stable than it is in the dichotomous model. Nevertheless, with substantial missing data in a matrix, research in the case of large $m$ is suggested along the following standard lines that have been used when $m$ is small. The first step is to check how closely, under various values of the parameters and number of components, the statistic does follow the chi-square distribution with the calculated degrees of freedom. If it follows it closely, then the statistic can be used immediately as a chi-square statistic with the associated probabilities. If it does not follow it closely, then computer programs that are used for analyzing data could incorporate the replicated simulation of responses according to the model and values of the parameter estimates for the data at hand; for example, 100 simulations could be run, and the value of the observed statistic could be compared with this simulated distribution. This research is not included in this article, but it is intended that the rationale provided in this article and the demonstrated potential for the application of it stimulates such research.

## A Global Fit Statistic for the Content Areas

A statistic to check the fit of the content areas also takes the form of a chi-square statistic. The latent distribution of profiles according to the parameter $\beta^*$ is divided into $G$ class intervals and the statistic

$$\chi_i^2 = \sum_{x=1}^{G} \frac{\left( \sum_{n \in g} x_{ni} - E[\sum_{n \in g} x_{ni}] \right)^2}{V[\sum_{n \in g} x_{ni}]} \tag{10}$$

is formed, with $\chi^2 = \sum_{i=1}^{I} \chi_i^2$ giving a global fit statistic on $(I-1)(G-1)$ degrees of freedom. As indicated earlier, the information from Equation 10 cannot be used to eliminate any of the content areas from profiles, but it can be used to help orient the profiles analysis.

## EXAMPLE

The data for the example are 577 profiles randomly selected from the 12,314 students who attempted the Tertiary Entrance Examinations in Western Australia in 1986. Because the example is presented in some detail for illustrative purposes, only 9 of the 28 available content areas are considered. The ASAT is used as a 10th content area, given that it is available for all persons, even though the procedure described would, as indicated earlier, obviate the need for a test such as ASAT to be used for purposes of scaling. Table 1 illustrates the raw data matrix showing the scores of 20 of the 577 profiles in the sample. The missing data in the matrix are a result of the students not taking those content areas.

### Parameter Estimates of the Content Areas

Table 2 shows the estimates of the four parameters of the content areas. The greater the value of the location estimate, the more difficult is the content area: Therefore, it is evident from Table 2 that physics, with the estimate $\hat{\delta}^*_p = 12.7$, is on the average the most difficult content area and English, with the estimate $\hat{\delta}^*_E = -9.7$, is the easiest. The order of the difficulties of the content areas is as expected from previous knowledge of their scaling according to the ASAT. ASAT itself appears relatively easy, although it has a relatively low discrimination. In addition, some of the more difficult content areas are contained in the sample. Perhaps an explanation of the parameter estimates for ASAT, which are similar to those of English, is that it, like English, needed to be targeted to the whole student population. The examinations in the other content areas are targeted more to the population of students who take that particular content area for examination. English is targeted to the whole population because at least one English content area needs to be studied, even if it is not included in the TES, and English literature is considered a specialist content area.

The estimates $\hat{\theta}^*$ of the dispersion parameters indicate that, for a given ability, the scores for English are the least spread, whereas for accounting, they are the most spread. This spread is qualified by the skewness and kurtosis estimates $\hat{\eta}^*$ and $\hat{\psi}^*$ and reflects the discrimination of the tests in the content areas. This can be seen

**TABLE 1**
**Illustration of a Raw Data Matrix**

| | Content Areas | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Profile ID* | *ASAT* | *English* | *English Literature* | *French* | *Accounting* | *Mathematics 1* | *Economics* | *Biology* | *Chemistry* | *Physics* |
| 1 | 71 | 51 | | | | 49 | 54 | 66 | | |
| 2 | 45 | 75 | 64 | 54 | | | | | | |
| 3 | 82 | 72 | | | | 82 | | | 60 | 61 |
| 4 | 80 | | 69 | 73 | | 85 | | | 85 | 68 |
| 5 | 50 | 43 | | | 18 | 41 | | | | |
| 6 | 67 | 64 | | | 60 | 64 | | 52 | | |
| 7 | 72 | 63 | | | | | 76 | | 64 | 65 |
| 8 | 67 | 71 | | 33 | | 62 | | 64 | | |
| 9 | 73 | | 58 | | | 75 | | | | |
| 10 | 70 | 46 | | | 58 | 80 | 39 | | | 41 |
| 11 | 28 | 54 | | | | | | | | |
| 12 | 53 | 22 | | | | | | | 58 | 60 |
| 13 | 77 | 57 | | | 88 | 77 | 74 | | 69 | 56 |
| 14 | 75 | 49 | | | 65 | 64 | 64 | | | |
| 15 | 56 | 72 | | 43 | 61 | 51 | | | | |
| 16 | 53 | 58 | | | 60 | 55 | 55 | 42 | | |
| 17 | 52 | 61 | | 87 | | 67 | | 63 | | |
| 18 | 62 | 47 | | | | 45 | 51 | 55 | | |
| 19 | 66 | 65 | | | | 56 | | | 21 | 42 |
| 20 | 48 | 59 | | | 75 | 73 | 58 | | | |

*Note.* ASAT = Australian Scholastic Aptitude Test.

TABLE 2
Parameter Estimates for 10 Content Areas

| Content Area | Location $\hat{\delta}^*$ | Dispersion $\hat{\theta}^*$ | Skewness $\hat{\eta}^*$ | Kurtosis $\hat{\psi}^*$ | Fit $\chi_i^2$ |
|---|---|---|---|---|---|
| ASAT | −8.9 | 61.5 | 3.3 | 7.0 | 7.5 |
| English | −9.7 | 65.3 | 4.9 | 5.5 | 12.9[a] |
| English Literature | 4.1 | 52.3 | 18.1 | −12.5 | 5.3 |
| French | 1.7 | 37.9 | −5.8 | −11.1 | 6.4 |
| Accounting | −1.7 | 37.4 | 25.0 | 2.7 | 4.1 |
| Mathematics 1 | −2.1 | 41.7 | 12.7 | −0.5 | 2.1 |
| Economics | −0.6 | 55.9 | 16.8 | 3.9 | 2.7 |
| Biology | −6.8 | 63.1 | 11.0 | 8.5 | 3.6 |
| Chemistry | 11.3 | 45.3 | 21.2 | 0.2 | 7.0 |
| Physics | 12.7 | 50.7 | 22.2 | 5.0 | 6.0 |

*Note*. ASAT = Australian Scholastic Aptitude Test.
[a] $\chi^2(36) = 57.60, p < .05$.

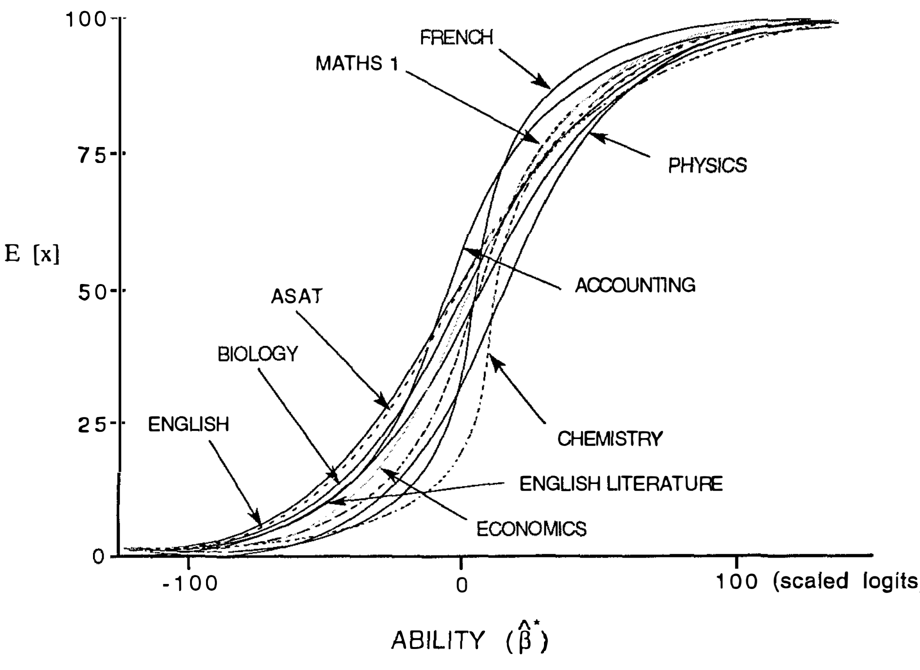


FIGURE 1 Expected value curve for 10 content areas.

from the estimated expected value curve (EVC) for each content area (analogous to the item characteristic curve [ICC] in latent trait theory for dichotomous items).

As indicated earlier, the expected value is given by $E[X_{ni}] = \sum_{x=0}^{m} x_{ni} \Pr\{x_{ni}\}$ and the EVC is formed by graphing these values as a function of the parameter $\beta^*$. Figure 1 shows the EVCs for the 10 content areas. In having different discriminations for the content areas, the EVCs of the ELM are analogous to those of the two-parameter logistic model (2PL) of latent trait theory (Birnbaum, 1968) for dichotomous items, in which the probability of a correct response is equal to the expected value. However, there is a major difference between the 2PL and the ELM: In the former, two parameters, the item difficulty and discrimination, are estimated from a response in one of two categories; in the latter, the minimum number of categories required to reveal different discriminations is three. In this case, different discriminations reflect different distances between the two thresholds on the continuum that mark off the three graded categories. With more than three categories, the discriminations reflect the distances between the thresholds. The dominant parameter that summarizes these distances is the dispersion parameter $\theta^*$, which is defined as the

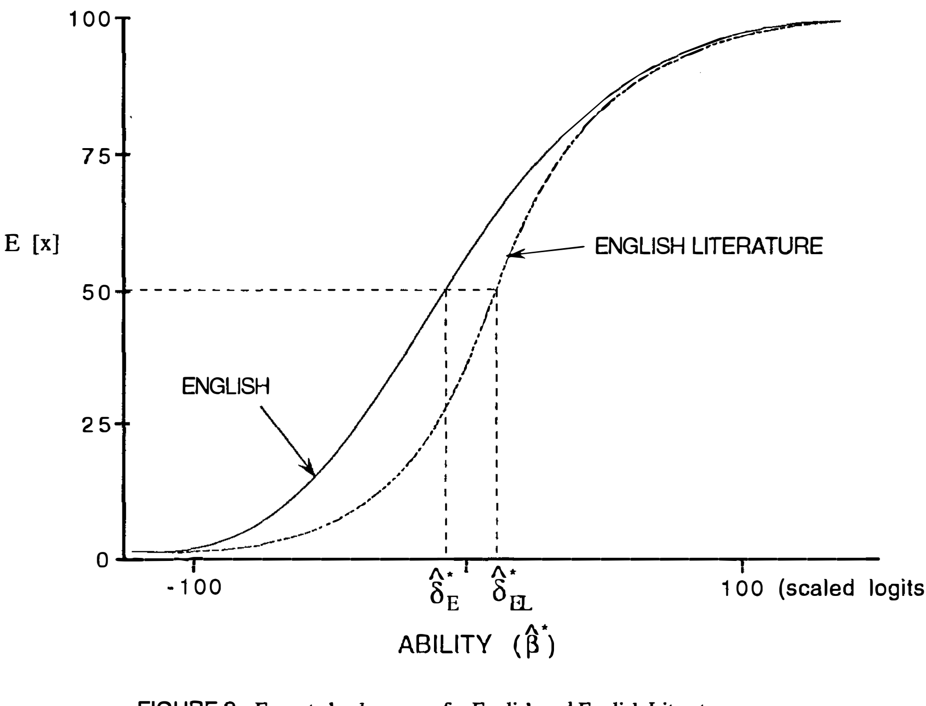


FIGURE 2 Expected value curve for English and English Literature.

average distance between successive thresholds (Andrich, 1985). If this parameter is small, then on the average, the thresholds are close together, and the rate of change of the expected score with respect to a change in the value of the location of the person on the latent variables is relatively rapid, that is, the discrimination is large.

To study more closely the effect of differences in parameter estimates, consider the EVCs for the pair of content areas English and English Literature shown in Figure 2. It can be seen that the EVC for English Literature is located to the right of the EVC for English essentially for the entire range of scores, showing that it is relatively more difficult to obtain the same observed score. That is, a student attempting English Literature would need to have a greater value on the latent trait than a person attempting English to obtain the same observed score. This is as expected because English Literature is a specialist area. It is clear that in order to amalgamate raw scores and not advantage or disadvantage students through their choice of content areas, it is necessary to take account of their relative difficulties and discriminations. This is carried out automatically by the estimate of $\beta^*$, thus compensating students who attempt content areas in which it is more difficult to obtain a high score. This effect is shown in the section on the analysis of profiles. Because the content areas discriminate differently, the effect is not constant. The

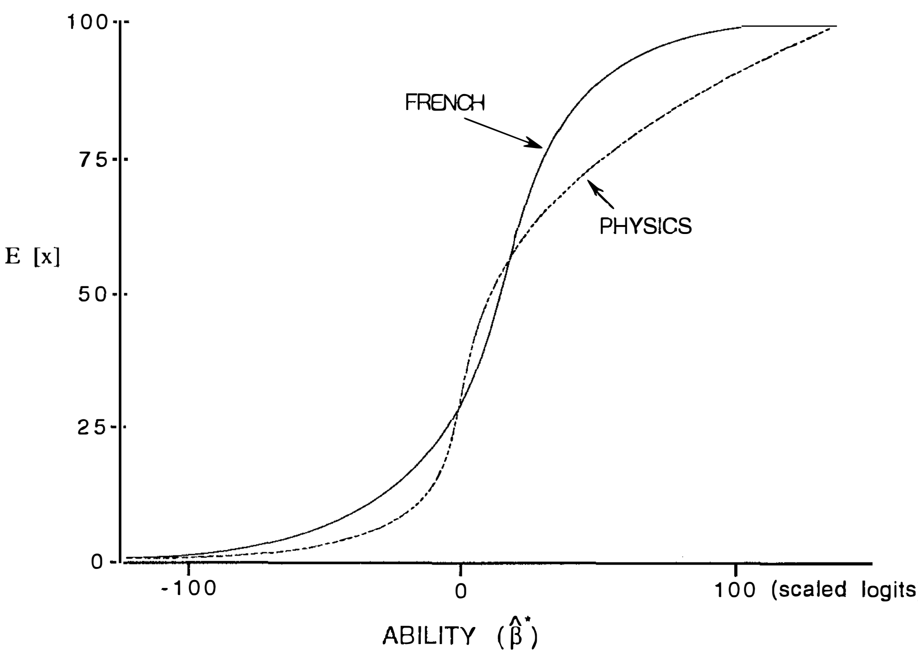


FIGURE 3 Expected value curve for French and Physics.

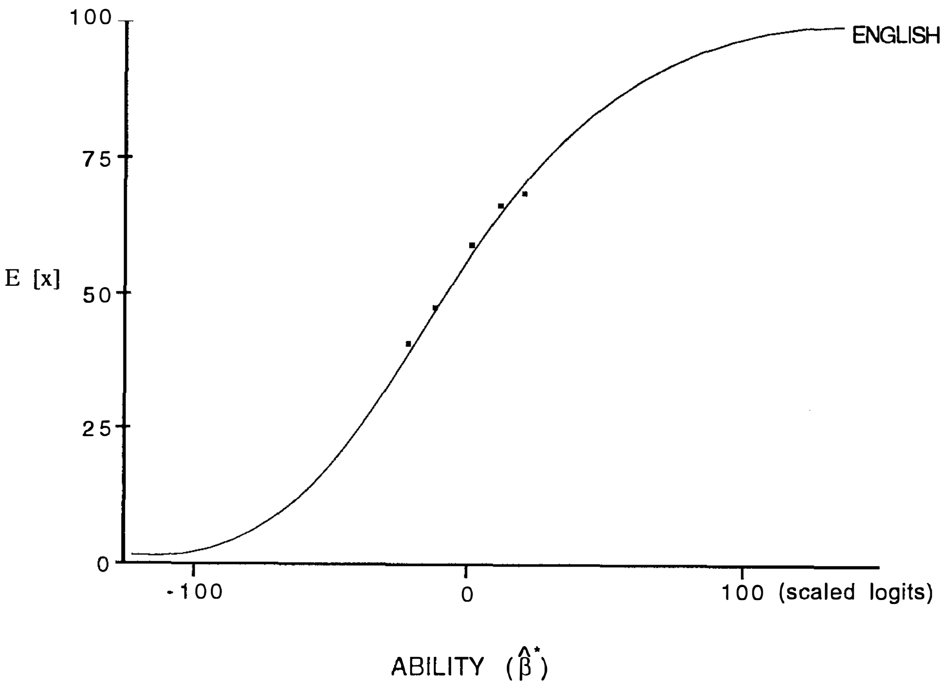


FIGURE 4 Expected value curve (with observed means) for English.

estimates of the parameters of the content areas and their role in estimating $\beta^*$ has a similar effect to the nonlinear traditional scaling of content areas described earlier.

To consolidate this point, consider the EVCs for French and Physics shown in Figure 3. The EVCs for French and Physics intersect at two observed scores, 25 and 60. Above 60, Physics is more difficult than French; between 25 and 60, French is a little more difficult than Physics. It can be seen from Table 1 that the estimates of $\theta^*$ for French and Physics are 37.9 and 50.7, respectively, confirming the graphical evidence that French discriminates more than Physics.

## General Fit of the Data to the Model

To orient the analysis of the profiles, a global test of fit according to Equation 10 is also shown in Table 2. It is evident that according to the global chi-square statistic, the data do not fit the model; however, it is also evident that this lack of fit is due to the content area English. In Figure 4, the EVC and the observed means for the class intervals of English is shown. The observed means in the class intervals, although statistically significant from the expected values, are nevertheless very

TABLE 3
Analysis of Profile 1

| *Content Areas* | *Expected Value* | *Observed Score* | *Standard Residual*[a] |
|---|---|---|---|
| ASAT | 58.6 | 71.0 | 1.3 |
| English | 59.0 | 51.0 | –0.9 |
| Mathematics 1 | 57.3 | 49.0 | –0.8 |
| Economics | 56.5 | 54.0 | –0.3 |
| Biology | 59.5 | 66.0 | 0.7 |

*Note.* ASAT = Australian Scholastic Aptitude Test. Latent ability estimate: $\hat{\beta} = -1.1$; Standard error $\hat{\sigma}_{\hat{\beta}} = 4.8$; $\chi^{*2}(4) = 4.17, p > .05$. $\chi_n^{*2}/f_n = 1.04$.
[a]$\chi^2 = 3.72$.

TABLE 4
Analysis of Profile 44

| *Content Areas* | *Expected Value* | *Observed Score* | *Standard Residual*[a] |
|---|---|---|---|
| ASAT | 56.8 | 61.0 | 0.4 |
| English | 57.4 | 78.0 | 2.3 |
| English Literature | 50.8 | 70.0 | 2.2 |
| Mathematics 1 | 55.2 | 5.0 | –4.8 |
| Economics | 54.7 | 61.0 | 0.7 |

*Note.* ASAT = Australian Scholastic Aptitude Test. Latent ability estimate: $\hat{\beta} = -3.0$; Standard error $\hat{\sigma}_{\hat{\beta}} = 4.7$; $\chi^{*2}(4) = 37.88, p < .05$; $\chi_n^{*2}/f_n = 9.47$.
[a]$\chi^2 = 33.82$.

close to them. From a substantive point of view, it is interesting and, perhaps, even encouraging that the one content area that shows least consistency is English. As indicated already, English or English Literature needed to be taken by all students, with the latter being a specialist content area. Thus, many students, including students from non-English speaking backgrounds, study this content area, which they perhaps would not do if they had total choice.

## Profile Analysis

To make the profile analysis concrete, Table 3 shows the profile for Student 1 of Table 1, the expected values for each content area according to the ELM, the standardized residual for each content area according to Equation 8, the latent ability estimate $\hat{\beta}^*$, the standard error of this estimate and the fit statistic of Equation 9. The fit statistic $(\chi_n^{*2} = 4.17, f_n = 4; \chi_n^{*2} / f_n = 1.04)$ suggests that Student 1 has performed relatively consistently across the content areas and that a single ability estimate for this student summarizes the profile.

Table 4 shows a relatively inconsistent profile of Student 44 with a relatively large fit statistic of $(\chi_n^{*2} = 37.78, f_n = 4; \chi_n^{*2} / f_n = 9.47)$. This student has performed very poorly in mathematics and much better in the English content areas. It is difficult to justify that a single score summarizes this profile.

Table 5 shows the profile of Student 51, which also has a relatively large fit statistic $(\chi_n^{*2} = 31.76, f_n = 4; \chi_n^{*2} / f_n = 7.94)$, but a pattern of strengths and weaknesses different from those in Profile 44. This student has performed well in science content areas but relatively poorly in English. The Economics result is consistent with the student's overall performance but is inconsistent with the poor score in English. It suggests that perhaps the examiners of Economics are marking for reference to content rather than for the expression of ideas. Similar profiles, particularly with the English and Economics inconsistency, were found with many students who are from non-English speaking backgrounds.

A fit statistic that is small indicates a consistent profile that accords with the model and that the single ability estimate summarizes the profile. Such a profile $(\chi_n^{*2} = 0.14, f_n = 5; \chi_n^{*2} / f_n = 0.02)$ is shown in Table 6.

Although the qualitative analysis of inconsistent profiles is considered as much a function of educational significance and professional judgment and availability

TABLE 5
Analysis of Profile 51

| *Content Areas* | *Expected Value* | *Observed Score* | *Standard Residual*[a] |
|---|---|---|---|
| ASAT | 66.8 | 59.0 | –0.9 |
| English | 66.4 | 29.0 | –4.3 |
| Economics | 64.2 | 68.0 | 0.4 |
| Chemistry | 56.7 | 75.0 | 1.9 |
| Physics | 55.9 | 79.0 | 2.3 |

*Note.* ASAT = Australian Scholastic Aptitude Test. Latent ability estimate $\hat{\beta} = 8.5$; Standard error $\hat{\sigma}_{\hat{\beta}} = 4.8$; $\chi^{*2}(4) = 31.76, p < .05$; $\chi_n^{*2}/f_n = 7.94$.
[a]$\chi^2 = 28.36$.

TABLE 6
Analysis of Profile 30

| *Content Areas* | *Expected Value* | *Observed Score* | *Standard Residual*[a] |
|---|---|---|---|
| ASAT | 60.5 | 62.0 | –0.2 |
| English | 60.7 | 59.0 | –0.2 |
| Mathematics 1 | 59.5 | 61.0 | 0.1 |
| Economics | 58.3 | 57.0 | –0.1 |
| Chemistry | 48.6 | 48.0 | –0.1 |
| Physics | 47.3 | 48.0 | 0.1 |

*Note.* ASAT = Australian Scholastic Aptitude Test. Latent ability estimate $\hat{\beta} = 1.0$; Standard error $\hat{\sigma}_{\hat{\beta}} = 4.0$; $\chi^{*2}(5) = 0.14, p < .05$; $\chi_n^{*2}/f_n = 0.02$.
[a]$\chi^2 = 0.12$.

of resources as it is of any statistical significance, taken at the $p < .05$ level for rejecting the fit of an individual profile, 47 of the 577 profiles (8%) in the sample warranted such an analysis. This indicates both a fair degree of conformity between the data and the model and perhaps a manageable number of profiles for a qualitative analysis.

This result gives a further opportunity to consolidate the perspective of this article—even though 8% of the profiles are rejected as fitting the model at the 5% level, this is not considered grounds for rejecting the ELM and seeking a more accommodating model. A more accommodating model, which perhaps has more parameters but which does not have the sufficiency property of the ELM, would absorb some of the inconsistency of profiles. Even though such a model might fit the data better, it would not serve our purpose of identifying inconsistent profiles worse—we are trying to identify inconsistent profiles efficiently and are not trying to model the data.

## Estimate of the Latent Ability

The estimates of latent ability takes account of the difficulty and discrimination of the content areas. Shown in Table 7 are the estimates for nine profiles. These are now considered more closely to illustrate the effects of taking different content areas on the estimate of the location $\beta^*$ on the latent continuum.

Profiles 559 and 269 have the same total score of 313 on the same content areas. Accordingly, they produce the same ability estimate. On the other hand, Profile 20 has the same total score of 313 from the same number of content areas (5), but instead of Chemistry and Physics, this profile contains Accounting and Mathematics 1. Because these content areas were easier than Chemistry and Physics, the ability estimate is substantially lower than for Profiles 559 and 269.

Profiles 129 and 529, again, have the same total score of 313 and the same content areas as each other, but this total score is based on only four content areas. Because it is much more difficult to obtain a total score of 313 from four than it is from five content areas, the ability estimate for these profiles is much greater (39.87 > 14.6) than for Profiles 559, 269 and 20. Thus, the ability estimate takes into account the number of content areas as well as their relative difficulty within the tertiary entrance set, which, as described earlier, permitted students to take three, four, or five such content areas.

Profiles of Persons 252 and 62, which have the same number of content areas, show the effective scaling through the model quite clearly: Profile 252 has the greater total score (322 > 311) but has a lower ability estimate (6.0 < 16.2) because Mathematics 1 and Accounting are easier than Chemistry and Physics.

Finally, Profiles 296 and 258 have the same content areas and have the same total score and, therefore, the same ability estimate, but the total score does not

**TABLE 7**
**Total Scores, Latent Ability Estimates, and Fit Statistics for Selected Profiles**

| ofile | Content Areas | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *ASAT* | *English* | *Eng. Lit.* | *French* | *Acctg.* | *Math. 1* | *Econ.* | *Biol.* | *Chem.* | *Physics* | *Total Score* $r$ | *Ability Estimate* $\hat{\beta}^*$ | *Standard Error* $\hat{\sigma}_{\hat{\beta}^*}$ | *Fit Statistic* $(\chi^{*2}/df)$ |
| 59 | 72 | 63 | | | | | 63 | | 51 | 64 | 313 | 14.6 | 6.3 | (2.35/4) = 0.59 |
| 69 | 73 | 61 | | | | | 57 | | 67 | 55 | 313 | 14.6 | 6.3 | (4.26/4) = 1.06 |
| 20 | 48 | 59 | | | 75 | 73 | 58 | | | | 313 | 3.8 | 6.3 | (5.04/4) = 1.26 |
| 29 | 79 | 82 | | | | 79 | 73 | | | | 313 | 39.8 | 7.9 | (0.95/3) = 0.32 |
| 29 | 77 | 79 | | | | 77 | 80 | | | | 313 | 39.8 | 7.9 | (0.53/3) = 0.18 |
| 96 | 69 | 37 | | | | | | | 62 | 48 | 216 | −0.1 | 6.5 | (13.76/3) = 4.59 |
| 58 | 56 | 53 | | | | | | | 54 | 53 | 216 | −0.1 | 6.5 | (2.1/3) = 0.70 |
| 52 | 56 | 72 | | | 71 | 60 | | 63 | | | 322 | 6.0 | 6.4 | (1.46/4) = 0.37 |
| 52 | 65 | 72 | 57 | | | | | | 58 | 59 | 311 | 16.2 | 6.5 | (0.784) = 0.20 |

*Note.* ASAT = Australian Scholastic Aptitude Test; Eng. Lit. = English Literature; Acctg. = Accounting; Math. 1 = Mathematics 1; Econ. = Econom-
Biol. = Biology; Chem. = Chemistry.

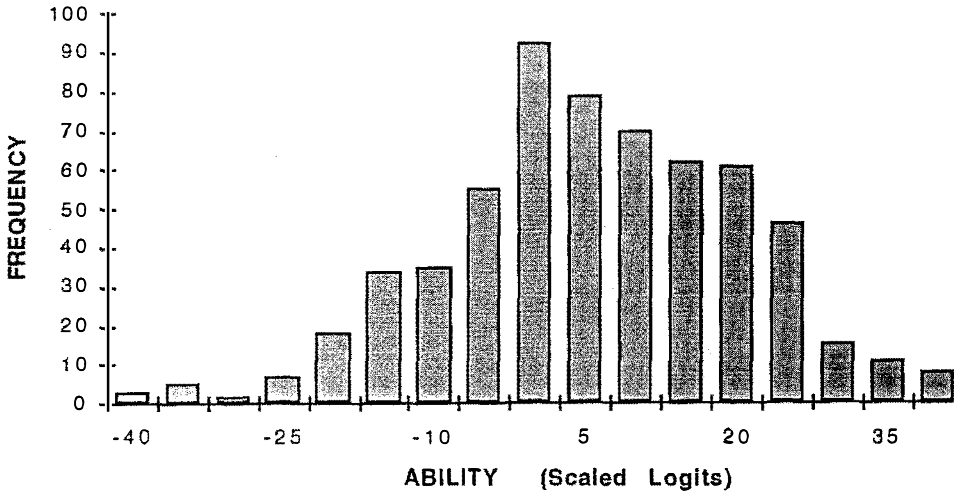


FIGURE 5 Distribution of abilities.

capture all of the information in the Profile 296 ($\chi_n^{*2} = 13.76, f_n = 3; \chi_n^{*2} / f_n = 4.59$), whereas it does for Profile 258 ($\chi_n^{*2} = 2.10, f_n = 3; \chi_n^{*2} / f_n = 0.70$). A single ability estimate for Profile 296 is difficult to justify.

Figure 5 shows the estimated ability distribution of the sample of persons used in the analysis. It is stressed that this distribution is observed and that no assumptions were made about the ability distribution. It appears to be a little negatively skewed.

## SUMMARY AND DISCUSSION

The previous analysis, which suggests that more than 90% of the profiles can be summarized by a single latent ability, confirms that, despite the variety of content areas studied by the students, their profiles are relatively consistent statistically. It was suggested that the choice of these content areas studied and the preparation for the examinations according to the requirements for entry contributes to the consistency. This is highlighted clearly in this example from Western Australia in which students had to take at least three content areas from the tertiary entrance set, including one from the humanities/social studies domain and one from the mathematics/physical sciences domain, and they had to study either English or English Literature. The expectation is that if students thought they were relatively weak in one of these domains, they would choose the content area within this domain with which they felt most comfortable and then would find support in their studies to overcome this weakness and make their profile more consistent than it might be otherwise. Of course, every student is not equally successful in this endeavor, but the tendency is evident. A further comment on English, which showed least conformity to the model, may be useful at this point. Although students had to take English or English Literature, if they had a content area in the humanities/social sciences in addition to English (in which nonspecialists tended to take this content area rather than English Literature), then they did not have to include the score in English in their TES. Thus, the procedure in Western Australia to some degree anticipated, although not for the statistical reasons uncovered in this article, the possible problem of including English in a single TES when many students were effectively forced to take English when they might have chosen otherwise with a completely free choice of content areas.

If a single latent ability does not summarize all of the statistical information in a profile, then it has been argued that the profile needs to be considered in relation to the intention of the selection and selection criteria at an individual level. This is especially relevant if a profile is in the margin of a cutoff score. For example, Profile 44 of Table 4, might be excluded if the single ability only was considered. However, the fit statistic suggests that the single latent ability does not summarize all of the information. The decision itself cannot be determined by the statistical model, which can only alert the decision makers to the inconsistency of the profile. In this example, an English department in a university might offer Student 44 a place (taking into account the profiles of other applicants), even though the single estimate of latent ability might indicate otherwise. Similarly, Student 51 in Table 5 might be permitted to proceed in areas requiring chemistry and physics, providing help was sought with English.

Although universities in Australia have been aware of the problems associated with using a single score for deciding entry, because of the lack of resources, they have been reluctant, as indicated earlier, to examine every applicant's profile. Our argument in this article is that it is possible to order the profiles according to the degree to which they require such examination. The number of profiles for special consideration can be reduced further if only those close to the cutoff point are considered. In deciding the region close to the cutoff in which inconsistent profiles might be studied more closely, the standard error of the latent ability estimate can be used. The number of profiles studied individually, once they have been ordered on inconsistency, depends largely on resources rather than on whether they are inconsistent at the 5% or 10% level of statistical significance. With large resources, perhaps 50% of the profiles could be examined individually. In this article, we used the conventional 5% level for efficiency in making specific points statistical inference.

There is, of course, in contexts similar to those of the example, concern for the public acceptance of selection procedures. Two basic issues emerge. First, that the same raw total score may lead to different decisions, and second, that even with the same latent estimated ability, there may be different decisions. This, in part, involves informing and justifying the procedure to the public and making it explicit. For example, in Western Australia, and in Australia in general, it is accepted that the raw scores in the content areas need to be scaled in order not to penalize or advantage students as a result of their choice of area of study. The second case may be more difficult, but these decisions are more likely to be made at the level of a program within a university rather than at the level of the university as a whole, and it could be indicated in advance that profiles within a certain range of the cutoff score for a university as whole will be examined more closely with respect to the program to which they have applied. It is the case now that, although a university may have a minimum TES for acceptance into the university, different programs (e.g., law, psychology, medicine) have different cutoff scores above this minimum, which is a function of the number and quality of applicants.

Current perceptions among writers and decision makers concerned with the selection of students into universities in Australia seem to be that selection based on a single score constructed from heterogeneous content areas and that based on the study of complete profiles are contradictory (cf. Masters & Beswick, 1986), and it is a view supported in the wider literature (e.g., Mathews, 1985). Our major argument in this article is that they can be made complementary with the application of a latent trait measurement model. Further research is required in a number of aspects of the procedure suggested, but it is considered that it is sufficiently developed and that the implications are sufficiently important that this further research could now proceed relatively straight forwardly.

## ACKNOWLEDGMENTS

This research was supported in part by Grant 03/20/198 from the Australian Research Council. The Secondary Education Authority of Western Australia gave permission for use of the data in the example.

## REFERENCES

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69–81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 357–374.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology 1985* (pp. 33–80). San Francisco: Jossey-Bass.

Andrich, D. (1988). *Rasch models for measurement* (Sage university papers series on quantitative applications in the social sciences). Beverly Hills, CA: Sage.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Reading, MA: Addison-Wesley.

Dunn, T. R. (1982). An empirical demonstration of bias in HSC examinations. *Australian Journal of Education, 26,* 190–203.

Everett, J. E., & Robins, J. (1991). Tertiary entrance predictors of first-year university performance. *Australian Journal of Education, 35,* 24–40.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46,* 59–77.

Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions, 8,* 360–361.

Ludlow, L. H. (1985). Person and population as psychometric concepts. *Educational and Psychological Measurement, 45,* 851–859.

Masters, G. N., & Beswick, D. (1986). *The construction of tertiary entrance scores: Principles and issues.* Melbourne, Australia: University of Melbourne, Centre for the Study of Higher Education.

Mathews, J. C. (1985). *Examinations: A commentary.* London: Allan & Unwin.

Matters, G. (Ed.). (1987). *Tertiary entrance in Queensland (Australia): A review.* Brisbane, Australia: Queensland Department of Education.

Mislevy, R. (1985). *Recent developments in the factor analysis of categorical variables.* Princeton, NJ: Educational Testing Service.

Muriaki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Pedler, P. (1987). *Accounting for psychometric dependence with a class of latent trait models.* Unpublished doctoral dissertation, University of Western Australia.

Phillips, S. E. (1986). The effects of deletion of misfitting persons on vertical equating via the Rasch model. *Journal of Educational Measurement, 23,* 107–118.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 321–334). Berkeley: University of California Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press. (Original work published 1960)

Ryan, J. (1980, April). *Examples of person fit analysis.* Paper presented at the meeting of the American Educational Research Association, Boston.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs, 34*(2, Whole No. 17).

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46,* 359–372.

Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement, 48,* 657–667.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529–554.

Viviani, N. (1990). *A review of tertiary entrance in Queensland (Australia).* Brisbane, Australia: Queensland Department of Education.

West, L. H. (1985). Differential prediction of first year university performance for students from different social backgrounds. *Australian Journal of Education, 29,* 175–187.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In B. S. Bloom (Ed.), *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85–101). Princeton, NJ: Educational Testing Service.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Wright, B. D., & Douglas, G. A. (1977). Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement, 37,* 47–60.

Wright, B. D., & Masters, G. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. (1979). *Best test design.* Chicago: MESA Press.

# APPENDIX

## Rationale for Calculating Approximate Degrees of Freedom for Each Profile

### Case of Complete Data

In order to provide the rationale for calculating the degrees of freedom for each profile, consider first the case of complete data for dichotomous items in which there are $N$ persons and $I$ items. In this case, the total number of independent data points $T_{ni} = NI$, and the number of independent item parameters estimated is $I - 1$ for the items and $N$ for the persons. Therefore, the number of degrees of freedom $f_T$ for the data as a whole is given by $f_T = NI - N - (I - 1) = (N - 1)(I - 1)$. To partition the degrees of freedom per profile, we first partition the total degrees of freedom to the degrees of freedom per data point: $f_{ni} = (f_T \div T_{ni}) = [(N - 1)(I - 1)] \div NI$.

Because each profile has $I$ data points (one per item), the degrees of freedom per profile are given by $f_{ni}(I) = (f_T \div T_{ni})(I) = [(N - 1)(I - 1)](I) \div NI = [(N - 1) \div N](I - 1)$. For large $N$, this is clearly close to $I - 1$, which is intuitively appealing.

Although the chi-square distribution is defined for nonintegral degrees of freedom, it is more conventional to consider integral degrees of freedom. Therefore, the observed $\chi_n^2$ for each profile $n$ is multiplied by $N \div (N - 1)$ to give an adjusted $\chi_n^{*2} = \chi_n^2(N/(N-1))$ on $f_n = I - 1$ degrees of freedom. The $\chi_n^{*2}$ are not totally independent, but they are more convenient to interpret.

### Case of Incomplete Data

The rationale for the case of incomplete data is as previously stated. In this case, the number of data points is given by $T_{ni} \sum_n I_n$, where $I_n$ is the number of content areas in the profile of person $n$. Let $P_i$ (4 in the example of the article) be the number of parameters estimated for each content area $i$, $i = 1, I$. Note that there is one constraint on the difficulty parameters $\sum_{i=1}^{I} \hat{\delta}_i = 0$. Therefore, the number of parameters estimated for the content areas as a set is $I(P_i) - 1$. Let $N$ be the number of persons—one parameter is estimated for each person. Therefore, the total number of parameters estimated is $N + I(P_i) - 1$.

The total degrees of freedom $f_T$ is the number of independent data points minus the number of parameters estimated: $f_T = T_{ni} - [N + I(P_i) - 1]$. Therefore, the degrees of freedom per profile with $I_n$ content areas is given by $\frac{f_T}{T_{ni}}(I_n)$. The adjusted

$\chi_n^{*2}$ on approximately $f_n = I_n - 1$ degrees of freedom is given by $\chi_n^{*2} = \left(\frac{I_n - 1}{I_n}\right)\left(\frac{T_{ni}}{f_T}\right)\chi_n^2$ for each profile $n$.

## Example

In the data set of this article, the total number of data points $T_{ni} = 2{,}162$, and the total degrees of freedom $f_T = T_{ni} - 10(P_i) - (N + 1) = 2{,}162 - 40 - 577 + 1 = 1{,}546$. Therefore, each $\chi_n^2$ is adjusted according to $\chi_n^{*2} = \left(\frac{I_n - 1}{I_n}\right)\left(\frac{2{,}162}{1{,}546}\right)\chi_n^2$, that is, $\chi_n^{*2} = \left(\frac{I_n - 1}{I_n}\right)(1.40)\chi_n^2$. In the case of profiles having three, four, five, and six content areas, the adjustments are $\chi_n^{*2} = 0.93(\chi_n^2)$, $\chi_n^{*2} = 1.05(\chi_n^2)$, $\chi_n^{*2} = 1.12(\chi_n^2)$, and $\chi_n^{*2} = 1.17(\chi_n^2)$, respectively. Clearly, these adjustments are relatively small. This in itself is reassuring.