Detecting and Interpreting Local Item Dependence Using a Family of Rasch Models

Mark Wilson
University of California, Berkeley

This paper describes a method for detecting and interpreting disturbances of the local independence assumption among items that share common stimulus material or other substantive features. Dichotomous and polytomous Rasch models are used in an example to analyze Structure of the Learning Outcome (SOLO) superitems and examine the results for local independence problems. The results indicate that some disturbances were present among particular subsets of the items. *Index terms: local independence, partial credit model, one-parameter logistic model, Rasch model, rating scale model.*

When item response theory (IRT) models are applied to test data, peculiarities of the measurement situation might cast doubt on whether the assumptions on which the analysis is based are completely justified. Hambleton and Swaminathan (1985) discussed the full set of these assumptions. This paper concerns only one: the local independence assumption. Broadly speaking, local independence describes a situation in which the statistical dependency between modeled observations is a function only of the model's parameters. In the case of IRT, the parameters are the person and item parameters. If the local independence assumption is not met, there is *local dependence* (Yen, 1984).

In particular, a common problem occurs when a test is composed of items that have the same or similar stimulus material for subsets of items or if different subsets of items share other features. In such cases, clearly a choice must be made between viewing the test as a set of base-level items or as a set of *subtests* each composed of a set of base-level items. Considering the test as a set of subtests has one advantage: If the shared features have a significant impact on the validity of the local independence assumption among the base-level items, a model at the subtest level may reduce that impact to insignificance. However, subtest models in general, and the two described below in particular, do not describe the behavior of base-level items, but rather describe parameterized subset scores. Thus, the basic unit of test construction, the item, becomes "invisible" at the subtest level, making item analysis more difficult in a subtest analysis. A method is needed to translate the results of the base-level into the subtest level so that it is possible to examine the effects of individual items and to diagnose violations of the assumption of local independence.

When item parameters are estimated in IRT, structural errors in the model may lead to less accurate person parameter estimation. Andrich (1985) discussed this problem, noting that several advantages exist in analyzing cases in which combined items can form subtests rather than base-level items. First, fewer parameters may be needed to define an appropriate psychometric model, leading to a more parsimonious representation (but note, as

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 4, December 1988, pp. 353–364

0146-6216/88/040353-12$1.85

above, the "disappearance" of items from the analysis that results). Second, the subtest is the more appropriate accounting unit in constructing item banks. Third, ignoring the expected dependence between items seems questionable at a common-sense level.

Using a subtest-level model also should result in superior measurement in terms of better fit statistics. To assess fit, a likelihood ratio test could be used that would compare the base-level analysis (with its assumption of item-level local independence) with a subtest-level analysis (with its assumption of subtest-level local independence). The likelihood ratio technique for this particular situation was described by Andrich (1985) and its application was studied statistically by Rost (1982). The method described below also compares the results of a base-level with a subtest-level analysis. Here, however, the emphasis is on comparing estimates rather than comparing fit. The results of the base-level analysis are manipulated under the assumption of local independence to produce the equivalents of the subtest-level results. These results can then be tested for statistically significant differences, which provides a test of whether the base-level or subtest-level local independence assumption is more appropriate. Of course, this issue could be addressed by the likelihood ratio test, but comparing the two sets of estimates gives a clear and immediate basis for interpreting a negative finding.

The approach taken here also differs from that of researchers such as van den Wollenberg (1982) and Yen (1984), who used chi-square tests of fit to explore local independence violations. The present approach focuses on a particular type of local independence problem and would not be suitable to the wide range of possibilities that their techniques can detect. Narrowing the focus, however, makes the results of the method more easily interpretable.

Dichotomous and Polytomous Rasch Models

The Simple Rasch Model

The simple Rasch model (SRM; Rasch, 1960/1980) for the analysis of test data provides a way to place persons and items on a scale with a clear probabilistic interpretation of distance on the scale. For a dichotomously-scored item j, with difficulty δ_j attempted by person i with ability β_i, the probability of a correct response, $y_{ij} = 1$, is modeled as

$$P(y_{ij} = 1) = \frac{\exp(\sigma_{ij})}{1+\exp(\sigma_{ij})} \quad (1)$$

(Wright & Stone, 1979), where

$$\sigma_{ij} = \beta_i - \delta_j \quad . \quad (2)$$

When combining the probabilities of L items to find the probability of a response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iL})$, local independence is assumed. That is, with a vector of item difficulties, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_L)$, the probability of response vector $\mathbf{y}_i$ is

$$P(\mathbf{y}_i|\beta_i,\boldsymbol{\delta}) = \prod_{j=1}^{L} P(y_{ij}) \quad . \quad (3)$$

Local independence means that the ability of the person and the difficulties of the questions must be considered in calculating the probability of each response pattern, but once those factors are included a simple multiplicative rule tells how to combine the probabilities. Local independence is disturbed if some subgroup of persons has a special relationship with some subgroup of items not encompassed by the relationship

$$\sigma_{ij} = \beta_i - \delta_j \quad , \quad (4)$$

or if a special dependency existed between items or persons beyond that indicated by their relative difficulties or abilities. For each score obtainable on the test, a *score characteristic curve* (SCC) can be defined as a function of the item difficulties. If T_x is the set of all vectors of *1*s and *0*s of length L whose elements add to x, $\mathbf{t}$ is such a vector and t_k is its kth element, then the assumption of local independence as given in Equation 3 allows calculation of the score characteristic curves for a given test as

$$\mathrm{SCC}(x|\delta_1, \ldots, \delta_m) = \sum_{\mathbf{t}\in T_x} \prod_{\mathbf{t}} P(y_k = t_k) \quad (x = 0, \ldots, \mathrm{L}) \quad , \quad (5)$$

where δ_k is the difficulty of the kth item, y_k is the response to the kth item, and $P(y_k = t_k)$ is given by Equation 1. Figure 1 is an example of the set of SCCs for a test composed of four items; the curves are labeled by the appropriate score.

Polytomous Rasch Models

The simple Rasch model can be extended to situations involving polytomous responses. Two will be used in this paper: the Rating Scale Model (SCALE; Andrich, 1978), and the Partial Credit Model (CREDIT; Masters, 1982).

The CREDIT model's basic observation is the number of steps that a person makes beyond the lowest performance level, or in a rating situation the number of steps that an object is placed above the lowest level. Note that the number of ordered levels in each item need not be constant across all items. Consequently, the basic parameter is the step difficulty within each item. For an item with $m + 1$ ordered levels from 0 to m, the probability of person i with ability β_i being observed in the nth category of item j (i.e., $y_{ij} = n$) is

$$P(y_{ij} = n) = \frac{\exp\left[\sum_{k=1}^{n}(\beta_i - \delta_{jk})\right]}{1 + \sum_{t=1}^{m}\exp\left[\sum_{k=1}^{t}(\beta_i - \delta_{jk})\right]} \quad (n = 1, 2, \ldots, m) \quad , \tag{6}$$

where δ_{jk} is the difficulty parameter for the kth step in the item, and

$$P(y_{ij} = 0) = \frac{1}{1 + \sum_{t=1}^{m}\exp\left[\sum_{k=1}^{t}(\beta_i - \delta_{jk})\right]} \quad . \tag{7}$$

The number of categories in an item, $m + 1$, may vary within a test. The local independence assumption used in CREDIT is that, conditional on step difficulties, the interaction between a person and an item is independent between items.

The SCALE model is similar to the CREDIT model, except that the pattern of step difficulties is held constant across all the items in a particular calibration. This results in a more parsimonious expression for the model:

$$P(y_{ij} = n) = \frac{\exp\left\{\sum_{k=1}^{n}[\beta_i - (\delta_j + \tau_k)]\right\}}{1 + \sum_{t=1}^{m}\exp\left\{\sum_{k=1}^{t}[\beta_i - (\delta_j + \tau_k)]\right\}} \quad (n = 1, 2, \ldots, m) \quad , \tag{8}$$

where δ_j is the difficulty parameter for the jth item and τ_k is the threshold parameter for the kth step, and

$$P(y_{ij} = 0) = \frac{1}{1 + \sum_{t=1}^{m}\exp\left\{\sum_{k=1}^{t}[\beta_i - (\delta_j + \tau_k)]\right\}} \quad . \tag{9}$$

In SCALE, the number of categories is a constant. One way to connect the two models more explicitly is to note that adding a j subscript to the threshold parameters in the above SCALE equations (i.e., allowing them to vary over items) makes them formally equivalent to the CREDIT equations. Step difficulties can be expressed as thresholds by setting

$$\delta_j = \sum_{k=1}^{m}\delta_{jk} \tag{10}$$

and

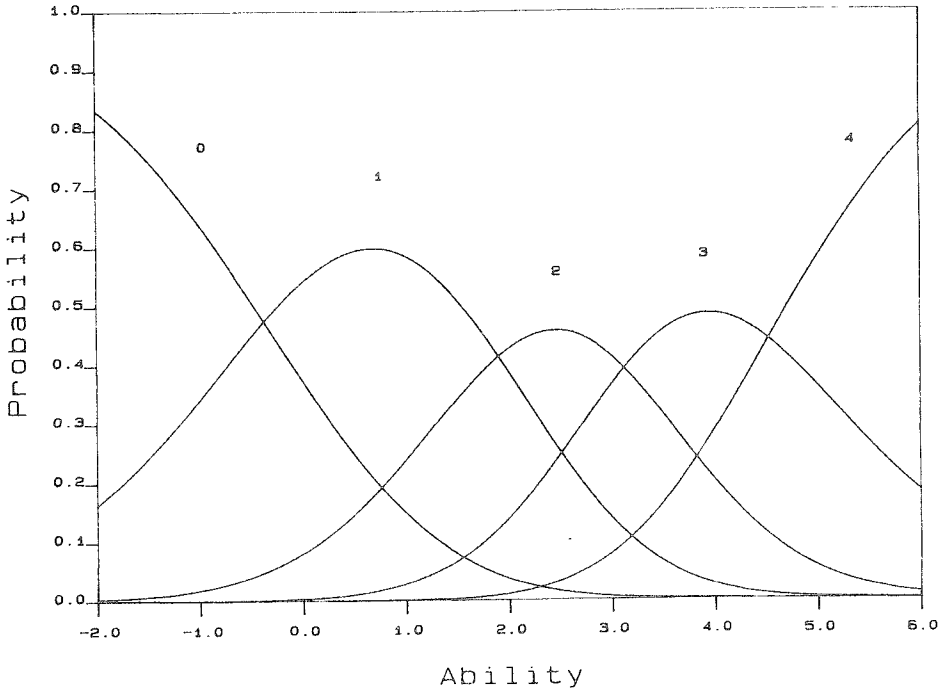
$$\tau_k = \delta_{jk} - \delta_j \quad . \tag{11}$$

The probability of particular categories of responses, given β, for both CREDIT and SCALE can be expressed as an item characteristic curve (ICC) simply by plotting the probabilities in Equations 5 through 8 against β. Then the step difficulties are represented in the ICC plots as the intersections of their respective curves. Figure 2 provides an example of the ICCs for a four-step item.

Andrich (1985) has described a useful alternative formulation of the SCALE model. He introduced a dispersion parameter into the SCALE model to replace the threshold parameters, naming the new model the Dispersion Location Model (DLM). Andrich showed that the DLM acts to constrain the threshold parameters to maintain a constant distance between them. The DLM is a specialized SCALE model that will be more parsimonious for items with more than three categories and identical for three-category items. The dispersion parameter θ represents the dispersion of responses: The greater the number of extreme responses, the smaller θ is, and vice versa. A slightly different technique connects the DLM with the SCALE and CREDIT models and eliminates the need to introduce another model and estimate its parameters.

Pedler (1988) has developed a way of estimating the dispersion parameter from the thresholds of the

Figure 1
Score Characteristic Curves (Synthetic ICCs) for Subtest 1



SCALE model:

$$\theta = \frac{\sum_{k=1}^{m}(m+1-2k)\tau_k}{\alpha} \qquad (12)$$

where

$$\alpha = \sum_{k=1}^{m}(m+1-2k)^2 \quad . \qquad (13)$$

This estimate will differ from a maximum likelihood estimate of θ based on the DLM model, but its interpretation will be the same in the situation described below. If s_k is the standard error for the kth threshold, then an approximate asymptotic variance of θ is given by

$$SE^2(\theta) = \sum_{k=1}^{m}\left[\frac{(m+1-2k)s_k}{\alpha}\right]^2 \qquad (14)$$

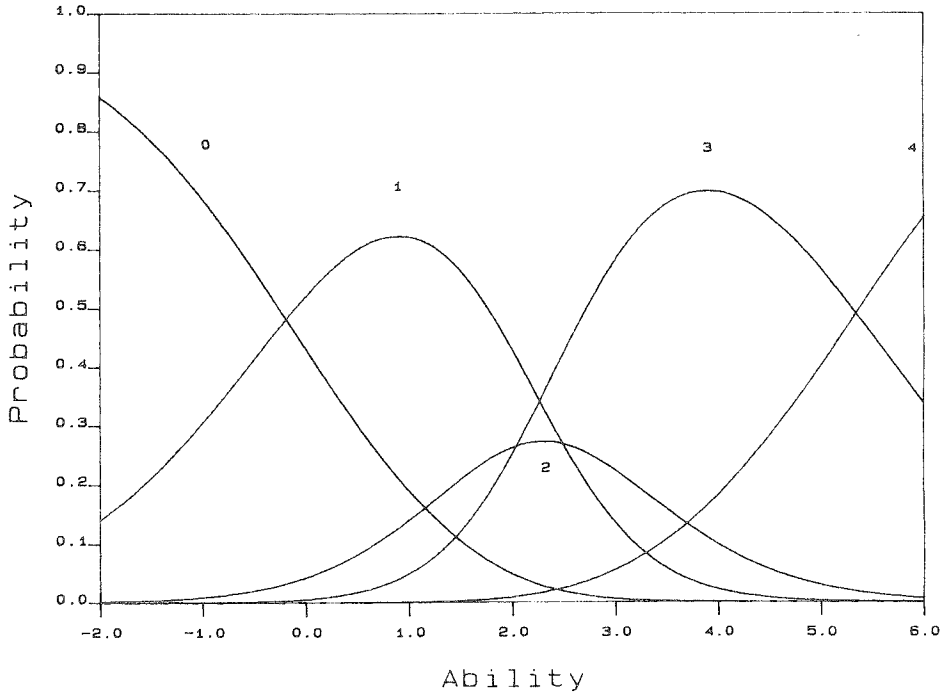
(Pedler, 1988).

Modeling Dependencies

One way to assess the impact of local dependence, especially among items sharing stimulus material, would be to calibrate the dichotomous data using the SRM, calibrate the polytomous data (scored by subtests) using one (or both) of the polytomous Rasch models, and then compare the levels of fit obtained. Andrich (1985) discussed this approach in the context of the DLM model. He demonstrated increased levels of fit associated with better ability estimation for the particular examples examined, but noted that available fit indices do not apply in cases with small numbers of persons and items.

The method described below not only focuses on deciding whether suspected violations of local independence have a statistically significant impact through tests of model fit, but also assesses violations of local independence by comparing dif-

Figure 2
Observed Item Characteristic Curves for Subtest 1



ferences in parameter estimates under various models. To simplify the explanation, base-level items are assumed to be dichotomously scored.

The method also attempts to aid interpretation of local independence violations by expressing those violations as disturbances in the expected pattern of estimations. To do this, the effects of the local independence assumption must be displayed in the metric on which interpretations are based: the logit ability scale. Going beyond statistically significant differences to interpretationally significant differences is often ignored in the literature because researchers believe that interpretationally significant differences must be embedded in an application. However, such an extended analysis is a crucial activity if measurement is to be seen by the wider community of educational and psychological researchers as something more than an obscure branch of statistics.

The effect of local dependence is assessed in the following manner. Consider a set of Rasch dichotomous item difficulties obtained from a calibration which may have been affected by some dependencies; further assume that the dependencies are reflected in a subtest structure. If there are m questions in each subtest, then the difficulties obtained from a simple Rasch calibration of all the questions (which assumes local independence between the items) can be compared with the step difficulties from a polytomous calibration (which assumes local independence only between the subtests). Note that for an m-item subtest, a score characteristic curve can be defined for the subtest, just as for the whole test. Then the equivalent of step difficulties will be given by the intersections of the consecutive SCCs:

$$\text{SCC}(x|d_{j1}, \ldots, d_{jm}) = \text{SCC}(x+1|d_{j1}, \ldots, d_{jm}) \quad (x=0, \ldots, m-1) \quad , \qquad (15)$$

where d_{jk} is the item difficulty for the kth item in the jth subtest. The following recursive formulas provide the equivalents of step difficulties (which will be called the "synthetic step difficulties" to distinguish them from the step difficulties obtained from a polytomous calibration) for a subtest using item difficulties from a dichotomous Rasch calibration:

$$\delta'_{j1} = -\log\left[\sum_{k=1}^{m}\exp(-d_{jk})\right] \quad (16)$$

and

$$\delta'_{jx} = -\sum_{k=1}^{x-1}\delta'_{jk} - \log\left[\sum_{t\in T_x}\prod_{t}\exp(-t_k d_{jk})\right] \quad (x = 2, \ldots, m) \quad , \quad (17)$$

where δ'_{jx} is the xth synthetic step difficulty in subtest j. These synthetic step difficulties can then be used to generate synthetic ICCs with the same intersections as the original SCCs. In fact, the curves generated are the same as the original SCCs. Thus, for the example in Figure 1, the curves representing both the observed SCCs and the synthetic ICCs are identical.

By using synthetic step difficulties, as expressed on the logit scale, the effect of assuming independence at the item level—rather than at the subtest level—can be ascertained. Generally, the effect can be seen by comparing the calibrated set of step difficulties, say for subtest j, $\delta_{j1}, \ldots, \delta_{jm}$, with the set of synthetic step difficulties, $\delta'_{j1}, \ldots, \delta'_{jm}$. A subtest with more dependency than predicted by the item difficulties would tend to have more of the logit scale occupied by the extreme scores (if $m = 4$, scores 0 and 4) and less occupied by the middle scores (scores 1, 2, and 3). This corresponds with the heuristic notion that, in general, dependence within a subtest will be expressed as an all-or-nothing response; either the student misunderstands the stimulus material and gets a 0, or the student comprehends it fully and gets a perfect score. Of course, this is only an extreme of the possible range.

Comparing Figure 1 with Figure 2 shows the importance of local independence assumptions. Here the same data were used to generate synthetic ICCs from a SRM calibration (Figure 1) and observed CREDIT ICCs (Figure 2). Clearly, different local independence assumptions resulted in varying patterns of estimates.

Andrich (1985) has proposed an alternative way to describe such dependencies using the dispersion parameter of the DLM model. Essentially, he noted that the greater the dependence, the greater the dispersion and the smaller the dispersion parameter. He calculated the dispersion parameter if the thresholds had arisen from dichotomous items under the local independence assumption using Equation 5.

If the observed dispersion parameter was less than the synthetic dispersion parameter, he interpreted it to mean that local dependence was detected by the DLM model. He also discussed how such problems can affect test reliability and how they relate to the attenuation paradox (Andrich, 1984; Andrich & Pedler, 1983).

Similarly, a synthetic θ' can be calculated from the synthetic threshold parameters and compared to the θ calculated from the threshold parameters estimated by a polytomous model. The indication of local dependence is the same as for the DLM. This description of the dispersion parameter has been expressed as a special case of the SCALE model, but as Andrich (1985) pointed out, extension to the CREDIT situation is quite trivial. In Equation 12, adding an item subscript to the dispersion and threshold parameters and standard error estimates allows calculation of a dispersion parameter for each item.

An Example

SOLO Subtests

The data used to illustrate the methods described above came from an application of the Structure of the Learning Outcome (SOLO) taxonomy (Biggs & Collis, 1982), which is a technique for classifying person responses according to the structure of the response elements. The taxonomy consists of five levels of response structure:

1. A *prestructural* response consists only of irrelevant information;

2. A *unistructural* response (U) includes only one relevant piece of information from the stimulus;
3. A *multistructural* response (M) includes several relevant pieces of information from the stimulus;
4. A *relational* response (R) integrates all relevant pieces of information from the stimulus; and
5. An *extended abstract* response (E) not only includes all relevant pieces of information, but extends the response to integrate relevant pieces of information not in the stimulus.

In a given topic area, persons are expected to move through each level from the prestructural to the extended abstract as their comprehension and maturity increase. Furthermore, the majority of responses should be classifiable into a level in the SOLO taxonomy that indicates the person's location on a latent dimension: "The structure of the SOLO taxonomy assumes a latent hierarchical and cumulative cognitive dimension" (Collis, 1983, p. 7).

In this example (Collis, 1986; Collis & Davey, 1984), a short piece of stimulus material (consisting of text, tables, or figures) was supplied and students were asked to answer several open-ended (but dichotomously-scored) questions concerning the material. Often, the stimulus material and the questions are referred to as a superitem (Romberg, Collis, Donovan, Buchanan, & Romberg, 1982; Romberg, Jurdak, Collis, & Buchanan, 1982); however, the more conventional term "subtest" will be used here.

The questions are linked to one of the higher four levels of the taxonomy. Responses are judged as acceptable or unacceptable according to an agreed set of criteria. This format uses the SOLO technique in reverse and results in four dichotomously-scored questions related to a stem, which can be used to indicate the level of concept development achieved.

A problem arises in interpreting response vectors that are not Guttman-like (Wilson, in press). In this example, five subtests in biology from an earlier study were used (Collis, 1986; Collis & Davey, 1984). The first subtest is shown in Figure 3. The first subtest was also the source of data for the examples in Figures 1 and 2. The data were gathered from 30 students in the 9th and 10th grades at a Tasmanian secondary school.

Analyses

The SRM and CREDIT analyses were performed on an IBM PC-AT microcomputer using the PC-CREDIT program (Masters & Wilson, 1988). The SCALE analysis was performed on a FACOM mainframe computer using the CREDIT program (Wright, Masters, & Ludlow, 1981). The programs report a lack of fit in terms of a transformed t statistic (Wright & Masters, 1982). According to these statistics, the SRM analysis displayed low levels of misfit for both persons and items, as did the CREDIT analysis. The SRM fit was also assessed with van den Wollenberg's Q statistic (1982), which was nonsignificant for each item and for the test as a whole. The SCALE analysis gave a large t statistic for Subtest 1.

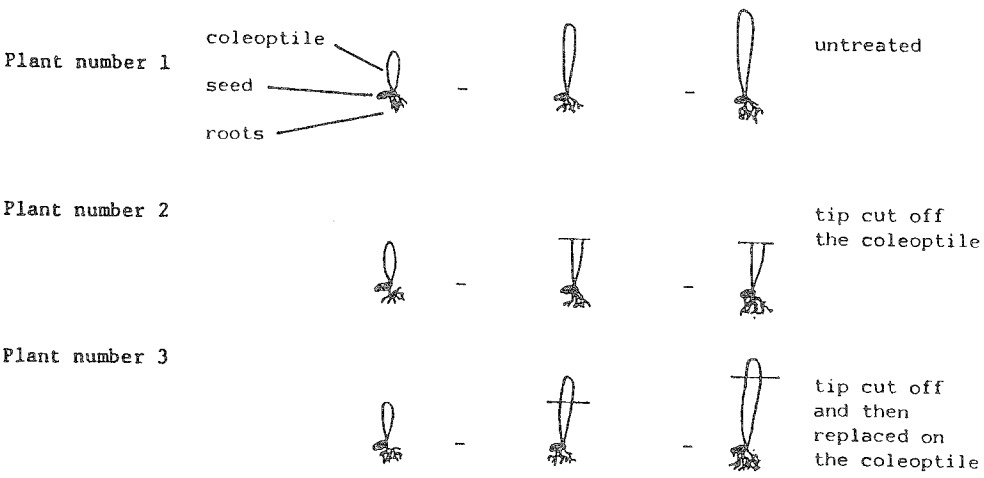
Results

Comparison of SRM and SCALE. The SCALE calibration was linearly transformed following the technique described in Hambleton and Swaminathan (1985) so that the location and spread of the student scores were the same as those for the dichotomous analysis. The synthetic subtest difficulties, the thresholds, and the dispersion for the biology subtests are shown in Table 1.

In order to conform most closely with the SCALE approach, the synthetic thresholds were calculated by determining the average distribution of thresholds for the four items so that each subtest had the same step pattern. The observed subtest difficulties, the thresholds (and their standard errors), and the dispersion for the SCALE calibration are also given in Table 1. They are all within two standard errors of the SCALE calibration equivalents. Clearly, the SCALE calibration provides no evidence that violations of the local independence assumption within subtests has affected the SRM calibration of the biology subtests. In other words, according to the SCALE calibration, dependencies between sub-

Figure 3
Subtest 1

A student performed an experiment in which he germinated three oat seeds and treated the coleoptiles in the following way.



Plant number	Height at Start	1 week	2 weeks
Plant number 1	1	2	2.5
Plant number 2	1	1.4	1.4
Plant number 3	1	2	2.5

U Which oat seedling had the tip cut off its coleoptile and not replaced?

M What is the height difference after two weeks between the seedling which had its tip removed but not replaced and the seedling which had its tip removed then replaced?

R How does the coleoptile tip affect the growth of a seedling?

E Develop a general theory which could have been tested by the above experiment, and list three other factors which would need to be controlled. (Use page opposite if not enough space below)

test items in the data are adequately summarized by the simple Rasch item difficulties, although individual items may need further attention.

Comparison of SRM and CREDIT. In the CREDIT analysis, it is not assumed that the steps follow the same relative pattern within each subtest, hence

Table 1
Synthetic and Observed Parameters From the SCALE Analysis

Parameter and Subtest	Synthetic	Observed Parameter	SE
Difficulty			
1	2.29	2.38	.36
2	2.98	3.36	.36
3	1.81	1.80	.36
4	1.89	1.63	.36
5	1.49	.98	.36
Threshold			
1	−3.39	−3.56	.68
2	−1.28	−1.33	.28
3	1.28	1.26	.28
4	3.44	3.63	.63
Dispersion	1.21	1.25	.08

each subtest can have a unique pattern of steps. The CREDIT calibration was also linearly transformed so that the locations of the student scores correspond to those for the SRM calibration.

The synthetic and observed dispersion parameters are presented in Table 2. The parameters show that only Subtest 1 has a notably smaller dispersion than in the SCALE model, lending weight to the suspicion that this item might be inconsistent with the local independence assumption. Table 3 provides more detail from the CREDIT results, and the synthetic ICCs and observed ICCs for this subtest are given in Figures 1 and 2, respectively.

Two differences are apparent between these CREDIT results and the SCALE results in Table 1. First, in Subtest 1, the second step is relatively more difficult than the third; this means that score 2 is never modeled to be the most probable response. In Subtest 3 neither score 1 nor score 3 becomes the most probable response at any point for similar reasons. This is different from the SCALE calibration, which indicated that each score in each subtest was most probable somewhere on the logit scale. Second, several subtests do not have SCCs for extreme scores. Subtests 2, 4, and 5 start at score 1 rather than 0; and Subtests 2 and 4 end at score 3 rather than 4. This occurred because of the absence of any responses for these score groups in the data. Although the SCALE model uses information in the other subtests to impute locations for

Table 2
Synthetic and Observed Dispersion Parameters From the CREDIT Analysis

Subtest	Synthetic	Observed Parameter	SE
1	.71	.35	.13
2	1.19	1.10	.63
3	1.51	1.48	.17
4	1.46	1.01	.44
5	1.29	1.42	.18

these scores, the CREDIT analysis reflects the limited range of the data in its results. The CREDIT analysis indicates that a broader range of student abilities is needed to fully analyze the subtests. Note that because of this problem with Subtests 2, 4, and 5, the dispersion parameters in Table 3 were calculated for maximum scores of 2, 2, and 3, respectively.

Table 3
Synthetic and Observed Step Difficulties From the CREDIT Analysis

Subtest and Step	Synthetic	Observed Parameter	SE
Subtest 1			
1	−.37	−.19	.88
2	1.90	2.49	.50
3	3.12	2.04	.51
4	4.52	5.34	1.22
Subtest 2			
1	−1.46	--	--
2	2.45	2.20	.46
3	4.63	5.44	1.17
4	6.28	--	--
Subtest 3			
1	−1.71	.39	1.23
2	0.00	−.44	.75
3	3.73	4.68	.68
4	5.21	2.82	.87
Subtest 4			
1	−1.71	--	--
2	−.02	−.02	.73
3	3.23	3.04	.50
4	6.06	--	--
Subtest 5			
1	−1.85	--	--
2	−.38	−.22	.87
3	2.16	2.57	.46
4	6.02	5.45	1.17

Clearly, the comparison between synthetic and observed steps confirms the problem with Subtest 1 noted by the dispersion comparison. This is most likely associated with the large t statistic for Subtest 1 in the SCALE analysis. In comparing Figure 1 with Figure 2, a modification seems to occur of the "all or none" dependency in the tendency for persons to get the M and R items either both correct or both incorrect. Inspection of the M and R items shows that they both involve calculating differences between "before" and "after" heights; for the M item the person must find the greatest difference, whereas for the R item the person must order the four treatments from smallest difference to greatest. These two tasks are quite similar and almost certainly not at different SOLO levels. This interpretation assumes that a score of 3 usually implies that the person answered the U, M, and R items correctly. The earlier SRM analysis demonstrated the orderliness of the SOLO items, so that assumption is reasonable. Similar reasoning supports the following comments on Subtest 3.

In Table 3, Subtest 3 reveals a problem that was not obvious from examining the dispersion parameter. Apparently the dispersion parameter is sensitive to cases in which dependence has affected the overall spread of the steps, but not to cases in which the overall spread is not greatly affected but the order expected under local independence is reversed (as indicated by the lower value for step 4 compared to step 3 in Subtest 3).

While dependence can be manifested in many ways, the dispersion parameter attempts to describe it using only one parameter, so that it is not surprising that it is not sensitive to this kind of problem. Subtest 3 suggests some dependency because the low expected probability of scores 1 and 3 indicates that students tend to answer the U and M items either both correct or both incorrect and the R and E items either both correct or incorrect. These response patterns indicate that items nominally at these two levels may actually be at the same difficulty level, although it would take a more detailed substantive analysis than is possible here to decide the issue.

Discussion

This paper has described a method for detecting violations of local independence among base-level items that are organized into subtests. In the example examined here, sets of items shared some feature that makes it likely that their behavior with respect to the latent trait will not be adequately modeled by a scheme that assumes local independence at the item level. This assumption will be a problem with the application of any item-level IRT model, and a technique similar to that described here could be developed for more complicated models than the Rasch model. The Rasch family of models is particularly suitable for this approach, however, as it has well-developed and interpretable polytomous extensions that embody the assumed dependence and that make inter-model comparisons relatively easy by having identical sufficient statistics for the person ability parameters. The discussion above assumed that the base-level items were dichotomous, which was correct for the example and makes the formulas simpler, but the restriction is not necessary.

Although examining relative fit may also be used to analyze this problem (Andrich, 1985), the present method not only allows a statistical comparison of the effects of the base-level local independence assumption, but also provides a straightforward means of describing exactly where the local independence assumption breaks down. While neither the t fit statistics nor the Q statistics indicated problems with Subtest 1 at the dichotomous level, the t statistic did in the SCALE analysis, but none indicated that there were problems with Subtest 3.

The technique consists of comparing parameter estimates obtained from the subtest-level analysis with "synthetic" parameters obtained under the assumption of local independence. This comparison can be simplified by constructing summaries of the parameters (the equivalents of Andrich's dispersion parameter).

The SOLO example demonstrated some features of the analysis: (1) the use of the SCALE model to examine overall or averaged local independence

effects and to deal with missing data; (2) the use of the CREDIT model to examine local independence effects more closely at the item level; (3) the potential for substantive interpretation of the effects; and (4) the use of the dispersion parameter as a summary statistic. The generalization of this technique to cases where the base-level items are polytomous is clear. A more comprehensive and widely-applicable strategy would be to construct an IRT model that allowed both subtest and item-level parameterization. An example of such a model has been given by Wilson (1985) where the subtest level is specialized to conform with a hypothesis of discontinuous development. More generalized models of this type are currently under development.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Andrich, D. (1984). *The attenuation paradox of traditional test theory as a breakdown of local independence in Person-item Response Theory*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Andrich, D. (1985). A latent-trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics*. Orlando FL: Academic Press.

Andrich, D., & Pedler, P. (1983). *The attenuation paradox in latent trait theory*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Collis, K. F. (1983). Development of a group test of mathematical understanding using superitem SOLO technique. *Journal of Science and Mathematics Education in South East Asia, 6*(1), 5–14.

Collis, K. F. (1986). A technique for evaluating skills in high school science. *Journal of Research in Science Teaching, 23,* 651–663.

Collis, K. F., & Davey, H. A. (1984). *The development of a set of SOLO items for high school science*. Hobart, Tasmania: University of Tasmania, Department of Education.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Masters, G. N., & Wilson, M. R. (1988). PC-CREDIT [Computer program]. Melbourne: University of Melbourne, Centre for the Study of Higher Education.

Pedler, P. (1988). *Accounting for psychometric dependence with a class of latent trait models*. Unpublished doctoral dissertation, University of Western Australia.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. [Expanded edition, University of Chicago Press, 1980.]

Romberg, T. A., Collis, K. F., Donovan, B. F., Buchanan, A. E., & Romberg, M. N. (1982). *The development of mathematical problem solving superitems*. (Report of NIE/ECS Item Development Project.) Madison: Wisconsin Center for Education Research.

Romberg, T. A., Jurdak, M. E., Collis, K. F., & Buchanan, A. E. (1982). *Construct validity of a set of mathematical superitems*. (Report of NIE/ECS Item Development Project.) Madison: Wisconsin Center for Education Research.

Rost, J. (1982). An unconditional maximum likelihood ratio for testing item homogeneity in the Rasch model. *Educational Research and Perspectives, 9,* 7–17.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123–140.

Wilson, M. (1985). *Measuring stages of growth*. Hawthorn, Australia: ACER.

Wilson, M. (in press). A comparison of deterministic and probabilistic approaches to measuring learning structures. *Australian Journal of Education*.

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

Wright, B. D., Masters, G. N., & Ludlow, L. H. (1981). CREDIT: A Rasch program for ordered categories [Computer program]. Chicago: University of Chicago, Department of Education, MESA Psychometric Laboratory.

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

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125–145.

Acknowledgments

The author thanks Professor Kevin Collis, University of

Tasmania, and Roy Pallett and H. A. Davey, both of the Education Department of Tasmania, for providing access to the data used in the example.

Author's Address

Send requests for reprints or further information to Mark Wilson, Graduate School of Education, University of California, Berkeley CA 94720, U.S.A.