JUST SUM THE MARKS: SPURIOUS WISDOM

Tim Dunne
University of Cape Town, South Africa
tim.dunne@uct.ac.za

Both formative and summative assessments may involve three-fold purposes: to rank performances of all participants as a norm-referencing task, to order test items by difficulty and make diagnostic inferences for intervention, and to adjudicate which respondents have attained performance of a specified criterion level (pass or distinction).The use of a weighted or unweighted total score (percentage) as reportable outcome is an unquestioned convention in statistics teaching and examinations, and elsewhere. Several issues arise: under what conditions is a sum convention adequate, and when is it deeply problematic? How can the test data disclose evidence against the use of the sum? What then can be done to offer paths to justifiable outcomes that can stand both ethical and legal scrutiny? The case is made for the importance of prior relative difficulty judgments recorded within the marking memo rubrics, and the use of appropriate statistical methods for ordinal data.

INTRODUCTION

Common practice in setting up statistics tests and examinations in school and higher education settings involves composing an instrument of a number of items. Items may be of several kinds: binary with zero-one score outcomes, multiple choice from say k options of which only one is fully correct, but k may vary from item to item, and constructed response items. We designate n as the number of candidates participating in a test.

MULTIPLE CHOICE

Each binary or multiple choice item (MCI) will generally have a maximum score one associated with a fully correct answer, and zero score attached to any and all incorrect answers (distractors) explicitly offered in the item structure. The incorrect answers are all scored as zero, even though there may be educational judgments which view some of the distractors as evidence of more severe errors than others within the same item.

If well designed for the purpose, the various distractors of any MCI may have inherent diagnostic value for future tasks of the tested cohort, or for teaching in subsequent cohorts. However, even where design was not so explicitly focused on diagnostics, there may well be post hoc information about test performance associated with the distractors of an MCI. This information can be accessed by comparing the overall test performances of the subsets of test candidates, as partitioned by the k choices of MCI, and doing so item by item.

In some applications concern is expressed about the possibility that random selection amongst k responses on an MCI will permit completely ignorant candidates to offer a correct response for a fraction $1/k$ of all repetitions. On this basis such a candidate will have an expected score (100% /k) on any MCI tests whose items all have k items, by any strategy of random selection.

The concern that ignorance should ever be rewarded with a positive score led to the mistaken adoption of negative score corrections, or the more sophisticated but confused notion of a guessing parameter, as in the so-called 3-parameter model in item response theory (IRT). As Andrich et al. (2012) note: it is not items that guess, but people. Better candidates make better guesses.

ITEMS WITH CONSTRUCTED RESPONSES

Each constructed response item is associated with a maximal point score, say m. The number m may vary across the test items. The point scores allocation within an item are associated with an initial marking memo. The explicit details of the memo are intended to specify how an appropriate item mark will be assigned to each student answer offered to the item. There are possibly n such answers, if each student attempts the item.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

The first objective of any item marking memo is to specify some relevant constructs and skills required from students to successfully respond to the item. The memo must reflect intersecting parts of the intended and the delivered curriculum of the course. It should not include material which is irrelevant to the curriculum, but may include unseen elements of relevance.

The second objective is to be fair to all students responding to the test item, in the sense of treating the response offerings by all test candidates alike. While the memo may be subjective, its specific detail must allow it to be applied in a verifiable manner, to each candidate.

In particular, the memo serves to guarantee a process by which each outcome mark for any person is replicable. The memo ensures that on remarking a person's script, the same mark will again emerge after any independent operation of the marking process based on the memo alone. This replicability is a necessary but not sufficient condition for item fairness.

Fairness also requires another condition: there should no feature in the item question itself that obscures what the memo is expecting. Similarly, there should be no feature missing from the item question whose absence misleads candidates into incorrect or incomplete answers. The item memo has to have a completeness by which it covers everything that a reasonable interpretation of language and structure of the item might interpret as being necessary. And nothing more than that.

The third objective is that the memo adequately describes a salient set of important distinctions that can be drawn between the candidates' responses to the item. We presume that the memo describes rules for an ordination of responses to any item in the test. It envisages a maximal number of possible groups to which we assign item responses, labelled by item scores 0, 1, 2, … *m*.

The fourth objective of each item memo is that there be no obvious duplication within the optimal answer for the item, of any material elsewhere in any other item in the same test. Each item is required to be probing some distinct elements of the knowledge outcomes required by the course. We seek to eliminate prospective double counting.

This fourth condition means the test is permitted to cover as broad a scoping of the course material as possible, given the time constraints imposed by the test or examination context. Inevitably no test can cover everything that is testable. We wish to ensure the test does not have narrow coverage.

The fifth objective is a resonance with the other items. This feature that joins all the items together is the expectation that they all relate to the same single type of proficiency that characterises the course being tested or examined. We expect that all items are positively associated. Association means that any two groups of candidates, one with a common high score on an item, and one with a common low score on the same item, will preserve higher and lower score profiles respectively, on every other item.

One consequence of the fifth objective of positive association between items is that the total score of any person is an intended indicator of proficiency. The question arises therefore, under what conditions will the total score verifiably be a valid and sufficient indicator of distinct proficiency levels. Note the sneakiness of language, the question asks about distinguishing between proficiencies, not adjudicating on any absolute level of competence. This intended reliance on the total score also constitutes a norm-referencing function of the test. In designing any test we are always setting ourselves a very complex agenda, just to achieve an adequate set of items.

Any criterion referencing has to rely on subject expert judgment external to the test construction, about the adequacy of levels of test performance for designations such as pass or distinction. These distinctions emerge from the ranking of items in terms of their relative difficulty, and values such as 50% or 75% as criterion levels are simply conventions of reportage, not inherently meaningful.

TIME ALLOCATION AND RELATIVE DIFFICULTY

Thus far the discussion of the item memo has also ignored the assumptions of the teacher about the necessary time interval a competent student might require to fluently articulate and record a constructed response to an item, so that the effort is of sufficient quality to warrant a maximal score *m*. This necessary time allocation for each item should be recorded with the memo. The total of these necessary intervals across all items justifies a minimal test duration.

The teacher will also have a view on the relative difficulty of the test items of all kinds. It is worthwhile to record these views. The outcome need not necessarily be a complete ranking of

the items, but it is useful to have as specific a sense of the ranking of items as teaching experience makes possible. At a minimum an ordination of clusters of items can be listed in a form that suggests their relative difficulty, or their relative demands for competence in the discipline or subject of the test.

This difficult task may be facilitated by cutting a test paper into page segments that present a single whole item, or where necessary, a single part of an item, together with the corresponding memo mark allocations. The page segments can then be sorted and ranked by the teacher's view of the associated challenge. Where several teachers are involved, it may be useful to preserve a record of both initial rankings of items by each teacher, and then gravitate to a consensus collective view of some kind.

These item time requirements and item difficulty rankings should be recorded against the items within the collective memo for the entire test. The point of such a resource is to have an originating view to contribute to the later data analysis. They also permit the post hoc exploration of performance against both norm and criterion referenced expectations, and considered decisions on test outcomes.

OBSERVED ITEM SCORES

The value of m to be assigned to an item may require astute insight into the kinds of responses that the teacher believes are likely to emerge from the candidates and context of the test. It is possible that post hoc evidence, from early grading results of an item, suggest changes to the memo may be required to cover the range of item responses and competence. There is likely to be diagnostic value in preserving both original and revised memos, and the explicit rationale for all changes.

These post hoc changes may in some cases include revising the notional maximal item score m. Such changes will require re-interpretation of the total scores attained by the candidates.

When the grading of a test item is complete, we are confronted with the range of observed item scores between 0 and m, and the corresponding frequency of each score category. For any item, any subset of student answers which are awarded a common item mark should be comparable with one another. They should all be robustly equivalent in quality, as encapsulated in the item memo. Such a subset of near equivalent answers should all be demonstrably superior to the all the answers assigned to any subset of item responses associated with a common lesser mark. The subset should simultaneously be demonstrably inferior to any answers for the same item in any other subset of student responses assigned a higher mark.

At the item level, the first type of anomaly that may arise from the context of using a test for a particular group of people, involves zero as the observed frequency of at least one of the possible item scores. The particular respondents are representative only of themselves and will generally not be any kind of sample, least of all a random sample. Thus it is possible even for a well-formulated item and its memo, that not all the mark categories 0, 1, … , m emerge in the marking process.

As teachers we may be delighted from an educational perspective with any zero frequency outcome for the lower range of scores within the memo requirements. Conversely, we are disappointed if the reduced range of observed scores features zero frequencies at the higher item scores.

If the number of candidates is small enough, these zero frequencies will occur naturally. But when the student cohort is large in number, these extreme-end anomalies require some analysis and perhaps merging and rescoring of item score groups. Similarly, there may be challenges to address if an item produces zero frequencies for one or several isolated mid-range score categories.

ORDINAL ASSOCIATIONS

The Gamma statistic for association in bivariate ordinal data (e.g. item scores, and by assumption, their various totals) is the fraction (P-Q)/(P+Q). It is obtained from the counts P and Q of all pairs of bivariate data cases (a, b) and (c, d) for which the orderings agree ($a < c$ and $b < d$) or disagree ($a < c$ but $b > d$). All pairs of cases with any tied categories are excluded. See Goodman and Kruskal (1954).

Gamma is a natural measure of association for ordinal variables, varying between -1 and +1, with zero implying no ordinal association. The variance of the Gamma statistics is a known function of P and Q. The Gamma value for any pair of ordinal variables is easily calculated in a spreadsheet from the corresponding two-way table of frequencies. It is useful to preserve the values of P and Q as well as Gamma.

The raw score data set will consist of k item scores for each of n persons. We allow let be m_i be the maximal possible score of the i^{th} item, and note $(m_i + 1)$ possible item score categories. We may summarise the pairwise relationships between any item and the total score by a two-way table of frequencies. If the maximum items score is m_i then the two-way table is $(m_i + 1) \times (\sum m_i + 1)$. The Gamma statistics and related counts are easily calculated

For a k-item test, with all items scores as 0/1, so that $m_i = 1$ and $\sum m_i = k$, these statistics can be estimated for all $k \times (k - 1)/2$ pairs of items, and well as for each of k items with both the test total of the entire test, and total of only the remaining $(k - 1)$ items. The Gamma results can be organized into a $k \times (k + 2)$ array.

Inspection of the leading square array within the table of estimates will permit judgments about the extent to which any pairs of items are effectively ordinal duplicates of one another ($Q \approx 0$), which item pairs are apparently unrelated ($P \approx Q$) or (alarmingly) negatively associated ($P < Q$).

The final two columns of the table reflect the strength of a positive ordinal relationship between item and total score. A strong positive relationship is inherent in the purpose of the test as a performance assessment covering the subject matter. This inspection will also confirm the extent to which each item resonates with the entire test (an inflated Gamma) and to which the association embodies the remainder of the test (from item scores paired with corresponding differences (total – item score) for each item, yielding unbiased Gamma's).

The same methods apply in principle to k-item tests with positive integer scoring ($m_i \geq 1$), in two ways. Firstly a similar array of Gamma statistics yields Gamma's for item pairs, and both inflated and unbiased Gamma's for items and test totals.

A simple device exists in two way tables to convert item × total score frequencies to the related frequencies for item × (total score – item score): namely discard first entry of second column and move cells up, first two entries of third column and move all cells up, and so on. The induced series of blanks at the bottom of the columns are interpreted as structural zero frequencies.

The consistency of discriminatory power of an item memo with maximum score m_i in relation to test performance can be examined by constructing Gamma arrays for m_i pairs of contiguous ordinal scores. The Gamma's for reduced tables will of course involve only the subsets of the n persons with appropriate item score categories. These simple methods using appropriate ordinal statistics permit insights into the extent to which the test scores admit an interpretation of coherent ordering of individual performances. They are predicated upon only the assumptions that the memo discriminates consistently on the basis of overall item performances.

CONCLUSIONS

A full Rasch model exploration of test data may not always be available. However the ordinal nature of test data will always require (non-parametric) ordinal approaches to assess the coherence of the test and its adequacy for purpose. This paper assembles key elements for a simple but thorough approach to test analysis, avoiding spurious use of parametric and regression constructs. This approach permits identification of anomalous item outcomes, memo changes, aggregation of adjacent item score categories, affirmation of coherent ordination of performances and expert judgment about attainment levels (e.g. pass and distinction). It will however be surpassed by further insights at item and test level offered by a Rasch model analysis.

REFERENCES

Andrich, D., Marais, I., & Humphry, S. (2012). Using a Theorem by Andersen and the Dichotomous Rasch Model to Assess the Presence of Random Guessing in Multiple Choice Items. *Journal of Educational and Behavioral Statistics*, *37*, 417-442.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, *49*(268): 732–764.