**Best practices in reporting analyses of questionnaires as objective rating scales of variable measures**

**What are the general objectives of this article?**

1. To clarify the meaning and implications of objective measurement applied in questionnaire analysis
2. To outline some recurrent problems that must be reported when constructing measures using questionnaires
3. To understand the importance of Wright maps in rating scale analysis
4. To encourage questionnaire analysts to be more explicit about techniques used in addressing data problems

**Abstract**

**Introduction:** Questionnaires are frequently used as rating scales of latent variables such as knowledge, anxiety and treatment outcomes. However, reporting the steps involved before generating the final ‘measures’ often fails to present known limitations and robust solutions to the problems common in questionnaire data.

**Aim**: To highlight some common problems in questionnaire data and suggest techniques of constructing objective measures during rating scale analysis.

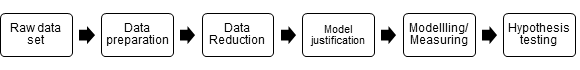
**Background:** The majority of questionnaire designers generate variable measures, either for educational or clinical research purposes, without providing adequate explanations of the steps taken to address inherent limitations that may worsen the error terms in the outcome measure. On the background that the usefulness of any measure depends on the least allowable error implies that best practice approach must be adopted during rating scale analysis. The best practice model states that the highest quality of scientific information in a discipline be engaged in addressing pertinent problems. Hence this paper proposes practical solutions to some shortcomings in reporting questionnaire analysis, based on modern theories of objective measurement in advanced statistics.

**Conclusion:** Cursory attention is given to the problems in questionnaire analysis as most users do not convincingly justify the applied measurement techniques before presenting variable estimation. Reporting the techniques used to address data complexity by engaging objective measurement parameters ensures best practice and emphasises the credibility of the outcome measure.

**Implications:** For a variable measure to be immediately useful, having limited error terms comparable to rating scales such as a clinical thermometer or height measure from a measuring tape, known limitations predisposing to increasing error must be identified and objectively resolved.

**Introduction**: Traditional methods of analysing questionnaires as rating scales in health and educational research “look good from afar” but may be far from being good, because rigorous treatment of questionnaire data proposed using advanced statistical methods is not always inculcated during the analysis (Boone and Noltemeyer, 2017; Colgrave, et al 2020). In spite of the popularity of questionnaires as valid rating scales of research variables, commensurate commitment to exemplary analysis is at best rudimentary, as applied measurement models are not adequately discussed nor justified (Bond and Fox, 2013). The primary objective of rating scale analysis is, foremost, to generate credible variable measure with minimal error. Attempting to achieve this purpose is, however, often confronted with common shortcomings found in questionnaire data and conventional analysis. Acknowledging that many misleading outcome measures from controversial computational methods are common, Leung et al (2012) suggested that there should be an appraisal framework for assessing quality of a rating scale and the outcome measure. Yet, Leung et al (2012) failed to incoporate the principles of objective measurement that may improve the rating scales or any measures produced using questionnaires in their suggestions. Consequently, Boone (2014) stated that routine publications about questionnaire analyses are neither consistent in reporting the key limitations nor provide adequate explanations on techniques used to minimise the error terms in the analysis. Bond and Fox (2013), extending the objective measurement models proposed by Wright and Masters (1982) into the human sciences, argued that some quantitative researchers add up raw scores from respondents and follow up by conducting parametric statistics as if questionnaire variables are measured on linear interval scales. Linacre (2021) noted that the problems found in analysing questionnaire data are not really new, but the lack of computer software and required skills are the major barriers to engaging modern theory. This argument, however, seems to have gained increasing attention, as more user-friendly computer software are now available for conducting objective measurement from questionnaire data, including Winsteps, RUMM2030+, ConQuest 5, Facets, WINMIRA and R (Rasch measurement analysis software directory, 2022). Simultaneously, the Institute for Objective Measurement (IOM) has stepped up formal online workshops or seminars, providing technical assistance and free access to software such as Bigsteps, Mplus, Minifac, Ministep and Openstat. Notwithstanding the advances in applied measurement research, questionnaire users in health-based disciplines are yet to adopt the most current techniques of variable measurement as the gold standard of rating scale analysis. Health care researchers (Melnyk, 2017; Hoyiso, et al 2018) are unequivocal in linking improved quality patient care with evidence-based practice: a model of treatment that advocates for objective research evidence in making clinical decisions. The implication here is that, for questionnaire analysis to represent efficient and effective outcomes conceptualised within evidence-based practice, exemplary statistical techniques must be engaged. On this background, this paper aims at two important objectives: (a) To explain the concept of objective measurement using questionnaires; (b) To argue that a rating scale analysis ought to follow and clearly report scientifically sound techniques rooted in objective measurement theory. Consequently, for clarity purposes, the use of a clinical thermometer (a routinely used measuring tool by nurses) is repeatedly introduced in this discussion to illustrate the meaning and application of objective measurement. The aim is that explaining objective measurement using clinically relevant measuring tool may motivate nurse-researchers (and other questionnaire users) to think deeper about rigorous techniques in analysing questionnaires as rating scale. The diagram below is a schematic of a possible step-by-step approach to using questionnaires as a rating scale presented in the ensuing article.

**Figure 1: The data analysis process**



**Understanding applied objective measurement**

**A good understanding of the meaning and practical applications of parameters of objective measurements is essential for generating objective measures from questionnaires, similar to using a clinical thermometer. From inception of the applied measurement theory, the founding fathers (Wright and Masters) were unequivocal on the four principles that define the philosophy of objective measurement of an identified variable. According to Wright and Masters (1982), objective measurement using questionnaire or any other measuring tool is underpinned by four principles: 1. The deliberate perception of a variable as a single entity or dimension (Unidimensionality); 2. The belief in the existence of a linear metric or possibility of modelling a variable into a linear scale (Linearity); 3. The belief that the process of measurement is so consistent that the same outcome (variable measure) will be reproduced without further modifications to the techniques or scale notwithstanding the subjects measured (Non-sample dependent measures); 4. The desire to compare a measured variable as higher or lesser among different groups. The first principle (unidimensionality) must be accounted for at the initial phase of developing the questionnaire** (Sakib et al 2020, Omolade et al 2022), **hence not discussed here. Principles 2and3 (linearity and non-sample dependent measurement) are the focus of this article contextualised within analysing questionnaires as rating scales. Principle 4 is about performing descriptive and inferential statistics on the measure generated after applying principles 2 and 3.**

**The four principles above mirror the core philosophy of scientific measurement applied in designing routinely used clinical tools such as clinical thermometer or measuring tape for grading patients’ heights. The implication is that if measures generated from questionnaires must gain comparable mathematical merits accorded to measures such as temperature, analysts must report the techniques engaged to adhere to all four parameters of objective measurement during the analysis. In other words, the ritual of merely counting numbers (from respondents) as a variable measure must be replaced with robust mathematical techniques rooted in objective assessment.**

**Measurement of a latent variable using questionnaires combines both relevant theories and applied mathematics** (Boone et al 2014, Boone and Noltemeyer 2017)**.** Bond and Fox (2014) stated that **while the theoretical inputs from literature evidence has grown, the background mathematics is stagnated by widespread simplistic approach to measure construction. For instance, to measure nurses’ evidence-based practice competence or ability, a number of items or indicators fitting into the recognised definition of evidence-based practice skills will be collated together. The usual practice in questionnaire design is to present the indicators as questions or statement of fact for respondents to endorse or disagree with. Yet, psychometricians recognise that, even at this preliminary stage of measuring, there can be misinterpretations of the indicators by respondents thus, the need for evaluating the psychometric properties of the questionnaire** (Boone 2016, Sakib et al 2020)**. Despite the excellent psychometric properties of a rating scale, estimating the outcome measure is still prone to remarkable computational errors unless the best techniques are engaged** (Leung, et al 2014)**. According to** Hilaliyah et al (2019), **error-prone measurement results by applying an observed score (X) as a variable measure without applying any techniques that minimise measurement (E) error from true score (T).** Boone and Noltemeyer (2017) **explained that observed score (X) is subjective being prone to contamination from respondents’ lack of concentration, guessing, deliberate random picking of answers or unavoidable distractions during data collection period. Conversely, engaging objective measurement techniques mean modelling questionnaire data to generate a measure with the least possible measurement error through application of mathematical techniques that factor in all possible sources of data contamination. Correspondingly, a visual display of linear continuum (principle 2) of the rating scale derived from the conjoint relationship among the indicators on a questionnaire cannot simply be assumed. More importantly, only at this point are the measures generated considered independent of the measured sample (principle 3); hence proving the consistency in the rating scale.** The importance of linearity (measurement of data on an interval scale) may be demonstrated with the example of a blood pressure monitor; because only by upholding a linear relationship of indicators can 1 mmHg be universally agreed as the true difference between a unit change along any point of the scale. This scale, like most interval measures, also possesses the property of ratio: for example a blood pressure reading of 120mmHg is exactly double that of a reading of 60 mmHg. Temperature, uniquely among common clinical measures is interval, but not ratio level, at least in its common form of units of degrees Celsius. A 1-unit change of temperature is a constant 1 °C anywhere on the temperature scale. However, 20 °C is not twice as hot as 10 °C. It is only ratio when measured on the Kelvin scale.

In contrast, scores estimated from a questionnaire lack either interval or ratio properties unless otherwise proven. Consequently, the difference in competency represented by two respondents who score, say, 20 points and 25 points on the questionnaire cannot represent the same difference as that observed between two respondents who score, say, 15 points and 20 points; or 10 points and 15 points, on that same scale. Further, a respondent whose observed score is 20 on an evidence-based competency questionnaire cannot necessarily be described as twice more competent than another respondent who scored 10. Similarly, the scores (25, 15 or 10) standing alone do not explain what competency the respondents possess based on the questionnaire initially administered. In response, **the steps outlined below will highlight the techniques that improve rating scale analysis in generating objective variable measure.**

Data Preparation

Data collected using questionnaires is analysed by converting a pool of categories and numbers into a simpler meaningful whole while adhering to the basic principles of applied mathematics. In a broad sense, variable measure from questionnaire bifurcates into data reduction and measurement modelling underpinned by objective mathematical theories (Boone et al 2014).  Applying useful models to manage the analysis ensures the whole procedure is scientifically robust and systematically objective. Yet measurement models, like any other useful concepts, cannot be applied capriciously but consequent upon screening the data for problems the applied model is meant to solve. Therefore, data examination is the first stage of analysing a quantitative data set. Data examination or inspection or data scrutiny or data preparation can be used synonymously to mean the initial screening of a data set for potential strength and limitations.

It is valuable to recollect that the observed numbers (data), from questionnaire administration, are products of unobservable interactions or reactions between the rating scale and the research variable (Wright and Masters 1982). Clinicians will agree that thermometer reading is an outcome of liquid mercury, coloured alcohol or other liquid in the evenly graduated glass tube reacting to a patient's body temperature. Underpinned by the same principle, a case can be made for questionnaires as measuring tools such that the numbers endorsed by respondents be treated as outcomes of reactions between the questionnaire items and the research variables investigated during data collection stage of a study. However, many problems can emerge at this stage demanding thorough investigation before committing to any meaningful calculations. For example, one may argue that some self-reported answers to the questionnaires do not represent the predicted reactions between the questionnaire and the research variable because the items on the questionnaire ought to be linearly related as in the case of graduated markings on the clinical thermometer. Thus, a fundamental question may be *Can there be responses in a questionnaire that do not match the predetermined measurement criteria?* This problem has been known by many researchers for many years yet glossed over or rationalised until the development of Rasch techniques (Boone 2016). Therefore, the critical function of a good measurement process is to examine the raw data for misfitting responses or related problems and report treatments applied (Boone et al 2014). Misfitting persons or responses do not abide by the objective measurement model, implying responses from such persons are inconsistent with the parameters of computing objective variables measures. Hence, the quality of data collected must be screened or inspected for anomalies that may limit the overall performance of the data. Some of the common problems and solutions are explained as follows.

**Coding errors:** Numerical codes or points or scores are allocated to the ordinal categories of a questionnaire (Bond and Fox, 2013). The majority of items are worded such that higher scores correspond to more positive responses. For example, 1, 2, 3 and 4 may correspond with Strongly disagree, Disagree, Agree and Strongly agree respectively on such positively phrased items in a Likert-type questionnaire. However, some items are worded such that lower scores correspond to more positive responses; such that, for example, 4 is allocated to Strongly disagree and 1 to Strongly agree. Boone et al (2014) noted that in most cases coding is completed at the initial development of the scale. But if required, recoding after data collection to correct any unintended mistake is possible when the indicators are screened or inspected.

**Completeness**: The completeness of the data collected highlights two issues. Foremost, all respondents may not use all the items provided on the questionnaire, hence the problem of missing data emerges. Secondly, the number of respondents may not match the sample size earlier calculated (sample size inadequacy) even though provision of 20% or more may have been made for attrition loss. Both routine and objective measurement analyses recognise these problems, but some researchers often fail to provide good argument on final decisions made. Addressing missing data in the classical test theory (CTT) is inconsistent; being unsupported by any measurement theory (Bond and Fox 2013). However, unlike CTT, the problem of missing data is well accounted for in the Rasch measurement technique because probability theory is used to assess observed questionnaire data against expected measurement model (Hilaliya et al 2019).

**The lack of linearity among the indicators:** Indicators’ conjoint order and linearity are fundamental requirements before adding scores together (Wright and Masters 1982). Until items are linked together based on the level of difficulty, calculations cannot begin because ordinal categories lack additivity and should not be mistaken as valid measures of a research variable (Boone 2016). Measures of a research variable become arbitrary labels if the items used in the measurement do not show meaningful relationships with other items on the same scale in such a way that one can decipher more or less of the research variable by marking the items or indicators (Wright and Masters 1982)

**“Perfect score” problems:** Perfect scores apply to respondents who choose the lowest or highest options for all the indicators, notwithstanding the coding system. Omolade et al (2022) argued that the primary aim of administering a survey is to discriminate among respondents but perfect scores do not align with this measurement requirement for a questionnaire to be treated as a valid rating scale. Wright and Masters (1982) had earlier explained that perfect scores indicate a mismatch between the survey items and the respondents. In other words, for the lowest perfect score, the survey or test may be too difficult and above the respondent's interest or ability. For the highest obtainable perfect score, the survey may be too easy, and the respondents' ability is above the difficulty presented by the indicators. Using the illustration of clinical thermometer again to illustrate a perfect score, the thermometer is calibrated within the range of human body temperature so that both lower and upper limits must never be reached. If a thermometer reads the lowest possible temperature gauge for any individual, then that reading requires further investigation and cannot be used as the basis for clinical intervention. The same problem applies if a thermometer records the highest possible temperature repeated times on a patient. Interestingly, Rasch measurement technique is designed to screen a data set for "perfect scores" and exclude such scores from the analysis because the misfitting scores cannot contribute meaningfully to objective measurement of a variable.

**Model fit:** Model fit suggests comparing observed data with the Rasch objective measurement model to assess fitting of items and respondents to the model. Boone et al (2014) called model fit diagnosis “data quality-control steps” to determine consistency in the observed scores. During questionnaire development, fit assessment focuses on the items or indicators while before outcome measure construction, respondents’ fitting is evaluated to diagnose individuals with complicated pattern of responses. Boone (2016) argued that model fit can identify increased measurement error from respondents overtly guessing answers, distracted when using the scale and any other pattern disconnected from predetermined measurement parameter. The implication is that respondents who clearly violated the fitting diagnostic criteria should be excluded from constructing the final measure of the research variable. Rasch theory of objective measurement uses indices of misfit to evaluate model fit (Bond and Fox 2013). The most frequently applied index of misfit as a rule of thumb is the mean square value of an outlier sensitive measure (Outfit) a value of this statistic above 1.3 indicates a poorly fitting model (Boone 2016).

**Measure construction (Wright map):**  Having identified some key problems that questionnaire data may present, any applied measurement model ought to provide mathematically sound solutions to the problem. In Rasch techniques, generating a variable measure is presented as a linear logit scale on a Wright map. The Rasch measurement models, which historically began with analysing dichotomous scales, have advanced into analysing polytomous Likert-type scales and the partial credit model of measurement (PCM) – scores awarded as partial credit of adjacent category (Bond and Fox 2013). Including the benefits that Rasch model produces interval score replica of the true score and outcome measure is estimated by probability, Wright map construction is a unique feature of Rasch analysis (Boone 2016).

Boone (2016) emphasised the importance of the visual representation of items, persons and measures on the Wright map so that respondents’ measures can be expressed in terms of skills or knowledge that respondents have mastered. For instance, presenting average scores (50% or 70%) of the evidence-based competency scale or any cognitive assessment is not sufficient for full inference; a relevant question is: what is the implication of the score on skills mastered? Correspondingly, Wright maps combine respondents, measures and items. Either for clinical reasons or educational purposes, having a visual qualitative interpretation of measures linked to questionnaire items helps to easily spot critical success factors in a population. The implication is that corresponding interventions can accurately target specific gaps in competency. Furthermore, Wright map analysis is powerful enough to describe levels of significance when group means are compared using analysis of variance (ANOVA) or t-testing. If there is a finding of significant difference between groups, a Wright map can isolate the indicators marking differences. Hilaliyah et al (2019) summarised the advantages of a Wright map into three: 1. Displays connections among measure, persons and indicators; 2. Indicators define respondent in a non-sample dependent version; 3. Respondents’ measure can be linked with demographics (age, years of practice, place of work) of interest.

Figure 1 is a Wright map of an unspecified variable depicting the linear relationship among persons and items based on the order of ability and difficulty respectively. The 12 items (Q36 to Q47) outlined on the right side are ordered from the easiest (Q41-bottom) to the hardest (Q42-top). This plot was derived from the output file on Winsteps following Rasch measurement modelling. The middle line is the “logit or Rasch” scale modified to present the measures from 0 to 7 logit; eliminating the infinite negativity and positivity measure. On the left side, each “#” represents 15 respondents and each “.” depicts 1 to 14 respondents. The mean measure for this group is 3.5 logit, and more than 420 respondents scored at least above the mean score for the group. If this Wright map were to describe, for example, the evidence-based practice competency of a group of clinicians, respondents located closer to the top of the linear continnum would be more competent than those closer to the bottom of the scale. In addition, any respondent scoring at least 3.5 (mean score) would find tasks Q44, Q43, Q38, Q36 and Q45 easier to perform than those below point marked “M" on the scale. Also, the map shows that more than average competency level to succeed in tasks Q46, Q39 is required; while tasks Q37 and Q42 will require even higher levels of competency. Through further examination of the respondents’ measure, researchers can also determine if those respondents who are expected to have low measures actually have lower measures than those respondents predicted to have higher competency. Since the Wright map presents the questionnaire items ordered according level of difficulty, Boone (2016) illustrated how the map can be used to produce a high-quality questionnaire for objective measurement. Further clarification on Rasch measurement models can be found in Wright and Stone (1979); Wright and Master (1982); Bond and Fox (2013) and Boone et al (2014).

**Figure 1 : Example of a Wright map of an unspecified variable**

|  |
| --- |
| MEASURE PERSON - MAP – ITEM (KNOWLEDGEandAWARENESS SCALE)  <more>|<rare>MOST DIFFICULT  7 . +  . |  | Q42  |  |  |  |  |  |  6 + Q37  |  . |  |  |  |  |  | Q39  |  5 +  .###### |  |  |  |  |  |  |  .########## |  4 . +  | Q46  |  |  . |  .############ **M|(MEAN OF PERSONS)**  |  | Q44  |  3 +  | Q43  .########### |  |  | Q38  |  |  ###### | Q36  |  2 +  |  |  .# |  |  |  | Q45  |  . |  1 +  | Q40  |  | Q47  |  . | Q41  |  |  |  0 +  <less>|<freq>EASIEST  EACH "#" IS 15: EACH "." IS 1 TO 14 |

**Selection of appropriate statistical software:** Using questionnaires is not as simple as it is sometimes purportedly reported, because the analysis process is confronted with complex problems demanding excellent knowledge in applied mathematics and computer programming to resolve. Factors that may influence the choice of Rasch software for constructing objective measure include accessibility, affordability or subscriptions, availability of manual, personal skills, past experience, training and certification. Finally, Table 1 below summarises the problems discussed in this article and suggested solutions.

**Table 1: Summary of techniques to rating scale analysis**

|  |
| --- |
| **Data analysis problems and treatment technique** |
| Coding errors: Inspect and recode data as appropriate |
| Completeness: Inspect data and check for attrition |
| Perfect scores: Apply the Rasch measurement model |
| Model fit: Apply the Rasch measurement model |
| Linearity: Apply the Rasch measurement model |
| Computing respondents’ measures: Wright map analysis |
| Deciding the best statistical methods: Follow good statistical guidelines |
| Choosing appropriate computer software: Winsteps and SPSS |

**Implication and conclusion:** Reporting the problems and techniques applied to resolve measure construction represent best practices in rating scale analysis. This implies analysing questionnaires as objective rating scales of research variables, adhering with the four parameters of objective measurement, including evidence of unidimensionality, linearity, consistency and additivity of the indicators on the questionnaire. Unfortunately, in many nursing journals, the use of questionnaires as a rating scale of research variable lags best technique proposed under theory of objective measurement. In this paper, objective measurement is conceptualised under the Rasch technique with special attention drawn to the use of Wright map. The argument is that questionnaire measures ought to pattern objective technique otherwise measurement error will derail the process into unintended misleading conclusions. Overall, this paper recommends to questionnaire users in nursing and related fields to embrace the rigours and benefits offered in applied objective measurement theory when reporting questionnaire analysis.

**Limitations**: Even though the aim of this paper is to ensure rating scale analysis is objectively conducted with least possible error, objective measurement using questionnaires ought to begin with the design of the measuring tool. Therefore, we suggest to questionnaire developers to study monographs, worked examples and relevant textbooks on developing high quality questionnaires in human sciences by authors such as Bond and Fox (2013); Boone (2016); Sakib et al (2020) and Omolade et al. (2022).

**Key points:**

1. Best practices embodied by advanced measurement technique are not routinely applied by researchers using questionnaires .
2. The steps outlined in this article should be reported during any rating scale analysis.
3. Engaging Rasch objective measurement techniques ensure variable measures are less prone to measurement errors.

# **References**

Bond T, Fox C (2013) *Applying the Rasch Model: Fundamental meausrement in the human sciences.* mawah: Psychology Press.

Boone JW, Noltemeyer A (2017) Rasch analysis: A primer for school pscychology researchers and practitioners. *Cogent Education*. doi:10.10180/2332286X.2017.1416898

Boone WJ (2016) Rasch analysis for instrument development: Why, When, and How? *Research Methods*. doi:10.1187/cbe.16-04-0148

Boone W, Staver J, Yale M (2014) *Rasch analysis in the human sciences.* Dordrecht: The Netherlands: Springer.

Colgrave J, Stasa H, Fraser J (2020) Validity and reliability of the psychometric properties of achild abuse questionnaire. *Nurse Researcher*. doi:10.7748/nr.2020.e1677

Hilaliyah H, Agustin Y, Setiawati et al (2019) Wright-map to investigate the actual abilities on math test of elementary students. *ournal of Physiscs*. doi:10.1088/1742-6596/1318/1/012067

Hoyiso D, Arega A, Markos T (2018) Evidence-based nursing practice and associated factors among nurses working in Jimma zone public hospitals, Southwest Ethiopia. *International Journal of Nursing and Midwifery*, 47-53. doi:10.5897/IJNM2017.0294

Leung K, Tarvena L, Waters D (2014) Systematic review of instruments for measuring nurses' knowledge, skills and attitudes for evidence-based practice. *Journal of Nursing*, 2181-2195.

Leung K, Travena L, Waters D (2012) Development of an appraisal tool to evaluate strenghth of an instrument or outcome measure. *Nurese Researcher, 20*, 13-19.

Linacre JM (2021) *A user's guide to WINSTEPS MINISTEP Rasch-Model Computer Programs Program Manual 4.80.0.* Chicago: IL:Winsteps.

Melnyk BM (2017) Models to guide implementation and sustatinability of evidence-based practice: A call to action for furthe use and research. *Worldviews on Evidence-based Nursing, 14*(4), pp. 255-256.

Omolade OK, Stephenson J, Padam et al (2022) Is this a good questionnaire? Dimensionality and category functioning of questionnaires used in nursing research. *Nurse Researcher*. doi:10.7748/nr.2022.e1842

*Rasch measurement analysis software directory* (2022) Retrieved from Institute for Objective Measurement: www.rasch.org

Sakib N, Bhuiyan AK, Hossain et al (2020) Psychometric validation of the Bangla fear of COVID-19 Scale: Confirmatory factor analysis and Rasch analysis. *International Journal of Mental Health and Addiction*. doi:10.1007/s11469-020-00289-x

Wright BD, Masters GN (1982) *Rating scale analysis.* Chicago: MESA Press.

Wright BD, Stone MH (1979) *Best Test Design.* Chicago: MESA Press.