Summer 2019, VOL. 32 NO. 3; ISSN 1051-0796

RMT

RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Overview of The Issue

In this issue of RMT, we have included Research Notes and several announcements that may be of interest to the Rasch measurement community.

The issue begins with a Research Note from William Fisher that relates measurement news from Sweden. In the second Research Note, A. Adrienne Walker provides a summary of the talk that she presented at the business meeting of the Rasch Special Interest Group (SIG) during the AERA conference. The third ResearchNote, from Agustin Tristan, focuses on separation statistics and alpha coefficients. Finally, Tsair-Wei Chien and Yang Shao provided a Research Note with an accompanying YouTube video that illustrates a procedure for simulating data in Winsteps.

After these notes, we have included announcements about several upcoming Rasch courses and workshops that may be of interest to the Rasch community, as well as a list of recent publications in *Journal of Applied Measurement.*

The issue concludes with minutes from the Rasch SIG business meeting and SIG announcements.

Finally, we created a survey to ask for your feedback on several ideas we have for RMT. Please complete the survey by August 1. Survey Link: http://bit.ly/RMTsurveyJune19

Please be on the lookout for more updates about the future of RMT!

Sincerely,

Your RMT Co-editors, Leigh and Stefanie

Rasch Measurement Transactions
www.rasch.org/rmt

News from Sweden's National Metrology Institute

The Research Institute of Sweden (RISE) is a National Metrology Institute responsible for maintaining and improving traceability to the International System of Units (often popularly referred to as the "metric system"). RISE recently proposed initiating and funding a new Center for Categorical-Based Measures (full disclosure: I was an advisor named on that proposal, and have ongoing collaborations with RISE). The content of the proposal focused on measures of short-term memory and attention span (Cano, et al, 2018a, 2018b, 2019) well known within the world of Rasch measurement due to the investigations of the Knox Cube Test conducted by Wright and Stone (1979; Stone, 2002).

The proposed project is historic in being the first effort by any metrology institute to date to initiate new unit definition, uncertainty, and quality assurance standards for psychological and social constructs. The viability and feasibility of these standards is presented in a body of work dating to the efforts of Thurstone, Rasch, Luce and Tukey, Wright, and others, with more explicit metrological formulations appearing over the last ten years or so (Mari & Wilson, 2014; Pendrill & Fisher, 2013, 2015; Pendrill, 2014; Pendrill & Petersson, 2016; Mari, et al., 2016; Maul, et al., 2018; Finkelstein, 2009; Fisher, 2009, 2012; Fisher & Stenner, 2016; Wilson & Fisher, 2016, 2018).

The funding application was denied. In response to reviews of the proposal provided by the Swedish VINNOVA innovation agency, the project team composed a reply that was published in a Swedish forum at https://www.dagensmedicin.se/artiklar/2019/05/03/jamforelser-kraver-kvalitetssakrad-matteknik/. In addition to RISE staff, the co-authors of this response include three longstanding Rasch measurement experts who were also members of the project team: Albert Westergren, Peter Hagell, and Curt Hagquist.

One of the team members from RISE, Jeanette Melin, posted an announcement of the publication on LinkedIn. The text was run through the Google Translate app to produce the translation shown below.

This area of measurement work will aspire to find a new professional society home in the upcoming International Measurement Confederation (IMEKO) Joint Symposium in St. Petersburg, Russia, 2-5 July. Cano, Melin, Fisher, Wilson, Andrich, Oon, Cavanagh, and a number of others planning to attend are excited about the inclusion of a fourth IMEKO Technical Committee, TC18 on Human Measurements, which will be a part of the Joint Symposium for the first time.

Given the success of the special session on psychometric metrology co-chaired by Wilson and Fisher at the IMEKO World Congress in Belfast, Ireland, last September, and of the 2016 IMEKO Joint Symposium they hosted at

UC Berkeley (Wilson & Fisher, 2016, 2018), national metrology institutes globally can be expected to take notice of Sweden's interest in establishing a new center for categorical-based measurements. IMEKO's TC18 on Human Measurements will likely provide a forum for reports on developments in this work undertaken by national metrology institutes around the world. Further information on this year's IMEKO Joint Symposium can be found at https://imeko19-spb.org/.

William P. Fisher, Jr.

References

Cano, S., Melin, J., Fisher, W. P., Jr., Stenner, A. J., Pendrill, L., & EMPIR NeuroMet 15HLT04 Consortium. (2018). Patient-centred cognition metrology. *Journal of Physics: Conference Series, 1065*, 072033.

Cano, S., Pendrill, L., Barbic, S., & Fisher, W. P., Jr. (2018). Patient-centred outcome metrology for healthcare decision-making. *Journal of Physics: Conference Series, 1044*, 012057.

Cano, S., Pendrill, L., Melin, J., & Fisher, W. P., Jr. (2019). Towards consensus measurement standards for patient-centered outcomes. *Measurement, 141*, 62-69.

Finkelstein, L. (2009). Widely-defined measurement--An analysis of challenges. *Measurement, 42*(9), 1270-1277.

Fisher, W. P., Jr. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement, 42*(9), 1278-1287.

Fisher, W. P., Jr. (2012). What the world needs now: A bold plan for new standards. *Standards Engineering, 64*(3), 1-5.

Fisher, W. P., Jr., & Stenner, A. J. (2016). Theory-based metrological traceability in education: A reading measurement network. *Measurement, 92,* 489-496.

Mari, L., Maul, A., Irribara, D. T., & Wilson, M. (2016). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement, 100*, 115-121.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement, 51,* 315-327.

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018). The quality of measurement results in terms of the structural features of the measurement process. *Measurement, 116*, 611-620.

Pendrill, L. (2014). Man as a measurement instrument [Special Feature]. *NCSLi Measure: The Journal of Measurement Science, 9*(4), 22-33.

Pendrill, L., & Fisher, W. P., Jr. (2013). Quantifying human response: Linking metrological and psychometric characterisations of man as a measurement instrument. *Journal of Physics Conference Series, 459*, 012057.

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement, 71,* 46-55.

Pendrill, L., & Petersson, N. (2016). Metrology of human-based and other qualitative measurements. *Measurement Science and Technology, 27*(9), 094003.

Stone, M. H. (2002). *Knox's cube test - revised.* Wood Dale: Stoelting.

Wilson, M. R. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement, 46*, 3766-3774.

Wilson, M., & Fisher, W. (2016). Preface: 2016 IMEKO TC1-TC7-TC13 Joint Symposium: Metrology across the Sciences: Wishful Thinking? *Journal of Physics Conference Series, 772*(1), 011001.

Wilson, M., & Fisher, W. (2018). Preface of special issue Metrology across the Sciences: Wishful Thinking? *Measurement, 127*, 577.

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

Google-Translated Article: "Comparisons Require Quality-assured Measurement Technology"

Note: Posted by Jeanette Melin of RISE, on LinkedIn on 3 May 2019. Translated from Swedish via Google Translate.

Without metrological references and traceability, it is not possible to make comparisons fully, write eight authors in a reply to Vårdanalys: *Claes Winzell*, unit manager at Rise measurement technology; *Leslie Pendrill*, researcher at Rise measurement technology; *Aslak Felin,* strategist at Rise measurement technology; *Jeanette Melin*, researcher at Rise measurement technology; *Evalill Nilsson*, researcher in the field of patient-reported measures in healthcare, Linköping University; *Albert Westergren*, Professor and Head of Research for the Research Platform for Health in Collaboration, Kristianstad University; *Peter Hagell*, Professor, Kristianstad University; *Curt Hagquist*, Professor and Director of the Center for Research on Child and Adolescent Mental Health, Karlstad University.

Published: 2019-05-03 07:00

This is opinion material. The opinions expressed here are the author's own.

Recently, the Agency for Health and Care Analysis presented its investigation work with proposals to the Government on how the national follow-up of health and medical care can be developed. A proposal that we see positively, partly in the form of how it has been worked out and partly in the form of the recommendations generated. With this post, however, we want to address a point that has not received space - the importance of quality-assured measurement technology in order to actually be able to make comparisons over time and the equality analyzes proposed, and these with a good reliability.

For several of the proposed indicators, it is concluded that the same question has been asked over time and with the same selection method, and that it would make comparisons possible. We agree that these are important preconditions for making comparisons, but we do mean that it is not possible to make comparisons fully, without metrological references and traceability.

Metrological references and traceability are two accepted concepts in measurement technology. This means that there are agreed and fixed references, for example, how heavy a kilo is or how long a meter is, to relate to measurements. Similarly, similar basic references would be needed to allow for reliability and comparability also with regard to the proposed indicators Health outcome of care, Person centering, and Accessibility to determine both whether unjustified differences prevail, and how care is developed and its equality aspects.

We also share Vårdanaly's views on the importance of not requiring the companies to collect and report more data, but to look at the possibilities within existing systems. A prerequisite for reliability and comparability is that the data that is collected is handled correctly and provides sufficient accuracy to follow development.

Here, too, there are important insights from the measurement technology to obtain, namely the central role that measurement uncertainties have.

In order to determine the reliability of the values of the various indicators, its measurement uncertainties need to be reported, that is, a range of spreads that the true value is within. This is a basic requirement for being able to say with certainty about a true effect, which does not depend on noise and uncertainty factors in the measurement.

This applies regardless of whether it is a person's blood pressure or a person's experience of how person-centered care is when you want to make a comparison of the value obtained in relation to, for example, a

previous measurement, measurements in different regions or specific target values.

In addition, as one of five points, in a speech in Dagens Medicin on March 29, 2019, Vårdanalys emphasizes that the national follow-up can be strengthened by developing the collection of patients' experiences and experiences of care. We fully share this perception, and are therefore critical of the fact that the assignment from the government is limited to indicators where data already exist today. With this you risk focusing on the low hanging fruits - if the Swedish health care system is to maintain a high international class, continue to develop and deliver an equal care, you must aim higher than that!

In 2018, Research institutes of Sweden, Rise, led a work on the need and opportunities to create a quality-assurance measurement technology infrastructure for categorical quantities within the healthcare system. After a series of bilateral meetings and a national workshop, we proposed a nationally coherent work - a Center for categorical measurements.

Such a center would have a corresponding role similar to our traditional national measurement sites and complement existing systems by screening among, developing new and implementing meaningful measures that enable reliable measurements and comparisons of effects of interventions, services and products, at both micro, meso and macro levels.

In the post on Dagens Medicin, representatives of Vårdanalys conclude with: “We see that the best way for this is to use our proposal as a starting point today. Please reconsider and change, but do not wait!” We therefore hope to highlight and clarify issues of quality-assured metrology as an important factor for the reliability of the follow-ups, which need to be considered in more depth in future work.

We also hope that RISE, the Research Institute of Sweden, as Sweden's Metrology Institute and with special expertise in handling self-assessments, will be included as one of all actors who must cooperate for the national follow-up of Swedish health and medical care and opportunities for a more equal care.

Rasch Measurement SIG Business Meeting Keynote Presentation Summary: *Person Fit of Individuals*

In practice, procedures that examine person fit are sometimes used by item analysts to examine the stability of estimated item parameters over population subgroups. However, the more common reason that an analyst may use person fit procedures that is found in the literature is to check the tenability of the inferences of a test score (regarding what a test taker knows and can do) given the test taker’s answers to the

individual items that are included on the test. For this note, these procedures are called individual person fit procedures. They represent quality checks that are important to do when test scores are used to make inferences about what individuals know and can do.

Individual person fit procedures seek to quantify how well the pattern of scored responses that a test taker gives matches with what is expected based on their total score (and the model used to generate that total score). For instance, for a test taker with an above average total score, it is reasonable to expect that they gave mostly correct answers to the easier items, mostly correct answers to the items of moderate difficulty, and gave mostly incorrect answers to the hardest items. If this response pattern is not observed, then the interpretability of the test score may be questionable. In this sense, individual person fit analyses provide information that can be used as a source of validity evidence based on response processes (AERA/APA/NCME, 2014).

In 2001, Meijer and Sijstma wrote of person fit analyses: "what is needed is an indication of how much misfit disturbs the estimated measures" (p. 130). This problem, how much misfit it takes to make the person's test score uninterpretable, still describes the state of practical application of person fit procedures in educational assessment contexts today in the United States. The purpose of the study discussed here is to illustrate an approach to identify when observed amounts of individual person misfit is too much for score interpretation and use. In other words, when is person fit no longer *good enough* for the score to be interpreted and used as an indicator of what the test taker knows and can do? The manuscript that describes the study and findings is currently under review, so the specific details of the rationale, procedures, and results have been omitted here. In summary, my co-author and I use real data from a criterion-referenced writing assessment that uses polytomous ratings to explore an approach to answer this question.

We establish a criterion for what *good enough* person fit looks like, and then we apply a procedure to detect when *good enough* person fit is not present. For practical reasons, we use the Rasch model and the Mokken Double Monotonicity (DM) model in the study. Because it is well known that the Rasch model and the Mokken Double Monotonicity (DM) model share many of the same theoretical requirements for measurement, and because previous researchers have argued that the Rasch model is a more stringent version of invariant measurement than the Mokken DM model (e.g., Engelhard, 2008; Wind, 2014), we argue that these models, when applied together, support a way to conceptualize (and operationalize) *good enough* person fit. Specifically, we rationalized that if a person's rating profile misfit the Rasch model, but still fit with the Mokken-DM model well enough, then the criterion of *good enough fit* is met. If

a person's rating profile misfit both the Rasch and Mokken frameworks, then the criterion of *good enough fit* is not met.

In general, our findings support that individual person fit information obtained from Mokken analyses can help researchers better understand person fit information obtained from Rasch analyses. This suggests that person fit information obtained from both approaches (and used in a complementary way) may help practitioners and researchers find potential solutions for the problematic practice of reporting and using test scores of persons who do not exhibit *good enough* person fit. We acknowledge that follow-up analyses, like interviews with misfitting test-takers or qualitative analyses their essays, may be needed to make the best decisions about what to do about replacing a misfitting test score with better achievement information, but we hope that this research encourages other researchers, practitioners, and test developers to think about how person fit procedures can be applied and used in assessment settings in practice.

A. Adrienne Walker

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: AERA.

Meijer R.R., Sitsma K. (2001). Person Fit Statistic - What is Their Purpose? Rasch Measurement Transactions, 15(2), 823.

Engelhard, G., Jr. (2008). Historical Perspectives on Invariant Measurement: Guttman, Rasch, and Mokken. Measurement, 6(3), 155–189. doi:10.1080/15366360802197792

Wind, S. A. (2014). Examining rating scales using Rasch and Mokken models for rater-mediated assessments. *Journal of Applied Measurement*, 15(2), 100-132.

Theoretical Separation and Alpha Values for Objective Tests

Cronbach alpha and Separation G are important values to assess the quality of a test in terms of internal consistency as a proxy to reliability. Both elements have been extensively studied and discussed in classical test theory documents and Rasch analysis related papers (for instance Andrich, 1982; Fisher, 1992; Linacre, 1995; Wright & Stone, 1999). In general, an accepted alpha value for a test is at least 0.8 and up to 0.95, but nobody wants to reach the highest (and quite impossible to obtain) value of 1.0. Separations above 2.0 are desirable in a test. The problem

with those reference values is that they depend on subjective criteria. In fact, it is not possible to define reasonable values for a test if population is homogeneous and their measures have a low standard deviation.

Cronbach alpha is a function of the number of items (N), the sum of the item variances and the persons' variance (in raw scores). Under certain conditions, it is possible to calculate theoretical alpha values for a given test. If the difficulties of the items have a uniform distribution according to Wright and Stone (1999), then the values of alpha and G will only depend on the variance or the standard deviation SD of the measures. These authors did not specify the minimum and maximum expected difficulties for a given test. The consideration for this paper is that difficulties may follow a theoretical uniform distribution of items with difficulties from -1.5 to +1.5 logits. This assumption, called the Test Design Line, (TDL) is useful for a wide variety of tests (see Tristan & Vidal, 2007). This work presents two nomograms calculated for low values of SD normalized as percentage, Figure 1 shows theoretical alpha values and Figure 2 shows theoretical separation G values.

A test of $N = 68$ items is applied to a population of 251 persons. The mean is 0.21 logits or 37.2 raw score and a $SD = 0.56$ logits or 7.3 items. For calibration, I have used Winsteps®.

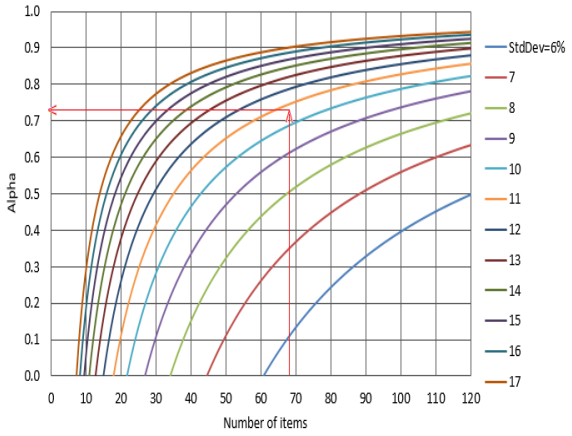


Figure 1. *Theoretical Cronbach alpha.*

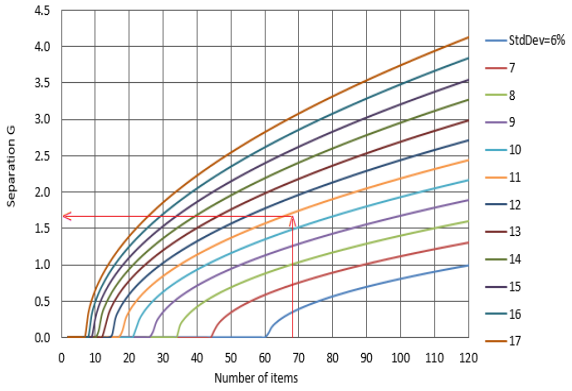


Figure 2. *Theoretical separation G.*

Example of Use

a) Which are the expected alpha and G?
I only need to calculate the normalized SD:
SD=7.3/68 × 100 = 10.7%
I can use the nomograms with the closest curve SD=11% or an interpolated curve for 10.7%. The red arrow shows that for 68 items the theoretical alpha is approximately 0.73 and the theoretical separation G is close to 1.70.

b) How close are those values to the empirical ones? We can see that the values provided by Winsteps® are

separation 1.7 to 1.76, reliability and Cronbach alpha between 0.74 and 0.75. Very close indeed!

```
TABLE 3.1 NATIONAL TEST VERSION:  1 C21 3.OUT Mar 26 14:16 2018
INPUT: 251 PERSONS  68 ITEMS  MEASURED: 251 PERSONS  68 ITEMS  2 CATS
-------------------------------------------------------------------------------
     SUMMARY OF 251 MEASURED PERSONS
+-------------------------------------------------------------------------------
|           RAW                          MODEL          INFIT        OUTFIT
|         SCORE     COUNT     MEASURE   ERROR      MNSQ   ZEMP   MNSQ   ZEMP
|-------------------------------------------------------------------------------
| MEAN      37.2      68.0         .21     .28      1.00     .0   1.00     .0
| S.D.       7.3       .0         .56     .01       .12    1.0    .19    1.0
| MAX.      57.0      68.0        1.95     .36      1.35    2.9   1.73    3.5
| MIN.      12.0      68.0       -1.83     .27       .75   -2.5    .62   -2.1
|-------------------------------------------------------------------------------
| REAL RMSE    .28  ADJ.SD     .48  SEPARATION  1.70  PERSON RELIABILITY  .74
|MODEL RMSE    .28  ADJ.SD     .48  SEPARATION  1.76  PERSON RELIABILITY  .76
| S.E. OF PERSON MEAN = .04
+-------------------------------------------------------------------------------
PERSON RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .75
```

The difference, if any, comes from a combination of various factors:

(1) Difficulties of the real test not strictly distributed from -1.5 to +1.5 logits

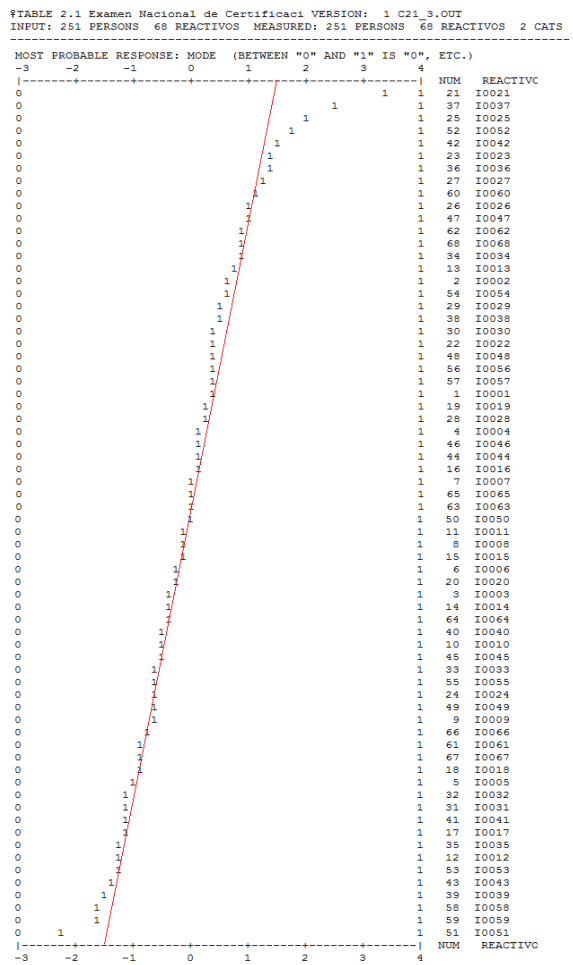
(2) Not all the items have good fit values

(3) Some items have low point biserial correlations,

(4) Answers with unexpected stochasticity inherent to the particular sample population (guessing, careless, idiosyncratic persons, and so forth)

Table 2.1 from Winsteps® provides the real distribution of the difficulties of the items, ordered by difficulty. The test design line (in red) corresponds to the uniform distribution from -1.5 to +1.5 logits. Some items do not fit the TDL (in particular in the extreme difficulties), but the approximation we obtain with the nomograms is remarkable.

Table 10.1 from Winsteps® shows the items ordered by misfit. About 10 items should be substituted or eliminated for future tests.




```
TABLE 10.1 NATIONAL TEST VERSION:   C21_3.OUT Mar 26 14:16 2018
INPUT: 251 PERSONS  68 ITEMS MEASURED: 251 PERSONS  68 ITEMS 2 CATS
-------------------------------------------------------------------------------
PERSON: REAL SEP.: 1.70  REL.: .74 ... ITEM: REAL SEP.: 6.76  REL.: .98
        ITEM STATISTICS:  MISFIT ORDER
+-------------------------------------------------------------------------------
|ENTRY   RAW                    MODEL|   INFIT  |  OUTFIT  |PTMEA|EXACT MATCH|
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZEMP|MNSQ  ZEMP|CORR.| OBS%  EXP% | ITEM
|-------------------------------------+----------+----------+-----+-----------+-----
|    37     27    251  2.4469  .2071|1.02   .1|1.30  1.2|A .06| 89.2  89.2| I0037
|    23     61    251  1.4228  .1513|1.10  1.1|1.21  1.6|B .03| 73.3  76.0| I0023
|    19    121    251   .2891  .1307|1.16  4.4|1.20  4.0|C-.04| 51.8  60.3| I0019
|    42     56    251  1.5405  .1557|1.07   .7|1.14  1.1|D .08| 77.7  77.8| I0042
|    52     50    251  1.6917  .1620|1.07   .7|1.14   .9|E .07| 78.9  80.1| I0052
|     1    119    251   .3233  .1308|1.09  2.4|1.12  2.4|F .09| 50.2  60.4| I0001
|    25     38    251  2.0397  .1798|1.04   .4|1.11   .6|G .10| 84.9  84.8| I0025
|    24    171    251  -.5976  .1396|1.02   .4|1.10  1.2|H .17| 66.5  69.0| I0024
|    50    141    251  -.0530  .1315|1.09  2.3|1.10  1.9|I .09| 52.6  61.1| I0050
|    30    113    251   .4261  .1312|1.08  2.1|1.09  1.7|J .11| 57.4  61.1| I0030
|    59    214    251 -1.6437  .1815|1.03   .3|1.08   .5|K .12| 84.9  85.3| I0059
|    61    182    251  -.8204  .1454| .99  -.1|1.06   .6|L .22| 72.9  72.9| I0061
```

Other Applications of the Nomogram

With the nomograms, it is possible to answer some questions, for instance:

[Question 1] Is it possible to get an alpha=0.8 and separation G=2.0 for a test

with 70 items, if the SD of the persons is 8%?

The answer is no, in this case alpha must be close to 0.52 and separation G close to 1.1.

[Question 2] I have a test with 50 items, alpha = 0.33 or a G = 0.75. How good is my test?

In general, people will say that it is evident that my test has a poor alpha, but the nomograms tell me that I must be careful before giving a general answer.

First of all, I must know the SD of the measures of the population.

What if *SD* = 8%? Figure 1 and 2 show that whit this *SD* the theoretical values are close to what I've got. I cannot expect a bigger separation or a bigger alpha.

What if SD = 15%? The nomograms show that I should expect G = 2.1 and alpha=0.82. In this case, I have to improve my test.

To have other values of alpha and G, it is necessary to modify the test design:

- Select new items from the item bank. They must have the same specifications (construct, content or taxonomic level) and better-fit parameters to the Rasch model.
- Eliminate of reduce noisy items (high misfit), they decrease the value of alpha and G.
- Check Guttman (muted) items. A set of items following the Guttman pattern may increase the variance of the measures, separation G and Cronbach alpha.
- Change redundant items, they increase the value of alpha and G. See the recommendations by Wright & Stone concerning the scale and the construction of measures.

Agustin Tristan

Instituto de Evaluacion e Ingenieria Avanzada

San Luis Potosi, Mexico

References

Andrich, D. (1982) An Index of Person Separation in Latent Trait Theory, the Traditional KR-20 Index, and the Guttman Scale Response Pattern. Education Research and Perspectives, 9:1, 95-104. https://www.rasch.org/erp7.htm

Fisher, W.P.Jr. (1992) *Reliability, Separation, Strata Statistics, Rasch Measurement Transactions, 6:3 p. 238.* https://www.rasch.org/rmt/rmt63i.htm

Fisher, W.P.Jr. (2008) The Cash Value of Reliability. Rasch Measurement Transactions, 22:1 p. 1160-3. https://www.rasch.org/rmt/rmt221i.htm

Linacre, J.M. (1995) *Reliability and separation nomograms. Rasch Measurement Transactions, 1995, 9:2 p.421.* https://www.rasch.org/rmt/rmt92a.htm

Tristan, L.A. y Vidal, U.R. (2007) *Linear model to assess the scale's validity of a test.* AERA 2007. ERIC ED501232. http://www.eric.ed.gov/PDFS/ED501232.pdf

Wright, B.D. & Stone, M.H. (1999) Measurement essentials. 2nd. Ed. Wide Range, Inc. Wilmington. https://www.rasch.org/measess/me-all.pdf

Rasch Simulation with Winsteps

The research questions are whether Rasch model fit statistics perform poorly when dimensions are interlaced (Tennant & Pallant, 2006) and what is the next step after removing Rasch misfit and cleaning the study data (Linacre, 2010).

Rasch simulation data show that unidimensional scales have the following features: (1) both Infit and Outfit mean square errors (MNSQ) are less than 1.5 (see the bottom-left box blots in Figure 1), and (2) Cronbach's α and dimension coefficient (Chien, 2012; Chien, Shao, & Jen, 2017) are both higher than 0.7 (see the top-left box blots in Figure 1).

If we assign one misfit item into the scale (i.e., a distinct domain correlation from 0 to 0.7 to the domain), the Outfit MNSQ (>1.5) can successfully identify another factor in existence, see the line of MNSQ = 1.5 at the bottom-left box plots in Figure 1. However, both Cronbach's α and dimension coefficient cannot precisely distinguish them on the occasion of one misfit item in a scale because all other scenarios are false negative, and the itmes still hold high values (>0.7) in their reliabilities and dimension tendencies (see the top-left box plots in Figure 1).

In contrast, if many misfit items are in data, only the dimension coefficient can be successfully applied to examine the scale quality (see the top-right panel in Figure 2) when fit statistics are substantially low (<1.5) showing false negative and all Cronbach's alphas are still high showing false negative also. Tennant and Pallant (2006) addressed that the Rasch model fit statistics performed poorly when dimensions were interlaced and the correlation between factors was near to 0.7. We verified that with lower domain correlations (e.g., from 0.3 to 0.0 interlaced with 33% or 50% proportion of misfit items, see Figure 2), the Rasch model fit statistics performed poorly.

Evidence related to dimensionality, such as dimensionality coefficients (see the unidimensional scenario in Figure 1) are necessary but not sufficient (see other scenarios in Figure 1) component of validity (Downing, 2003; Feldt, Brennan, 1989). The decision rule we propose is that **the next step after removing Rasch misfit and cleaning data should be to further report the dimension coefficient again** to ensure that there are not any interlaced items in existence.

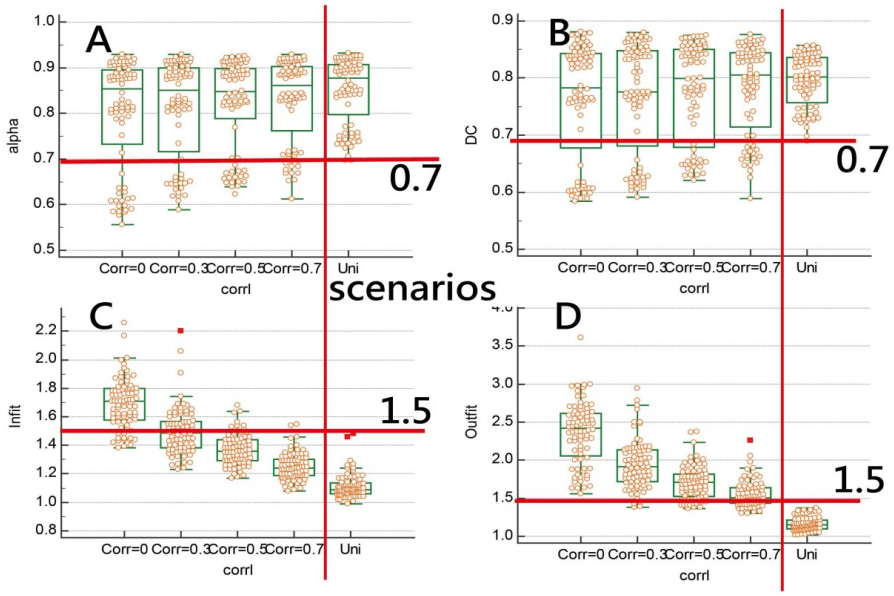


Figure 1. *Unidimensional scale and others with one misfit item correlated to the true score from 0 to 0.70*

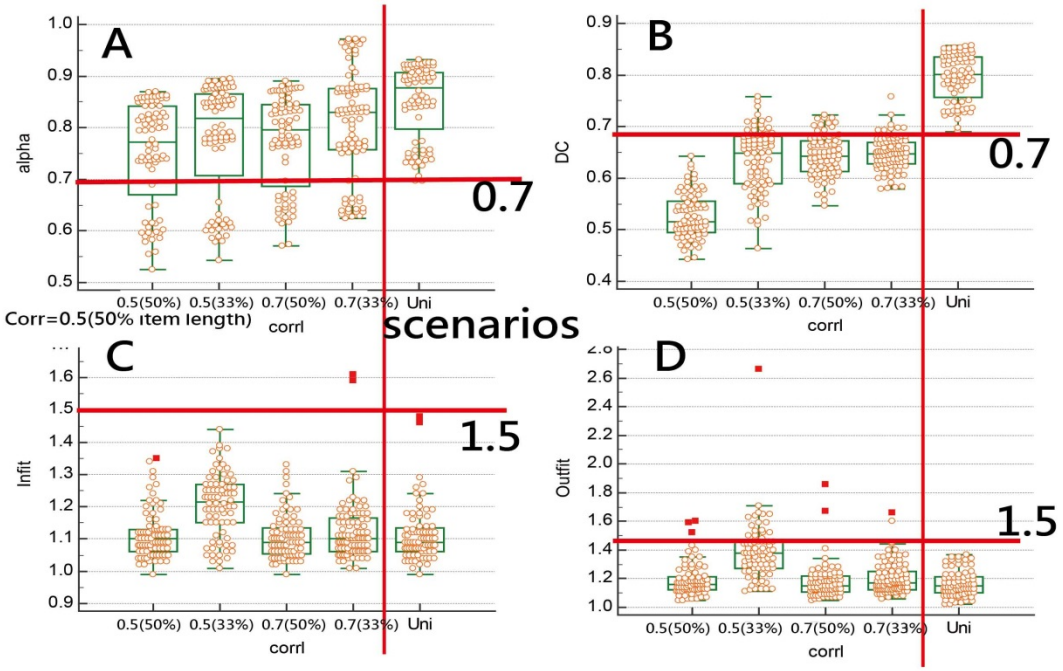


Figure 2. *Unidimensional scale and others with many misfit items correlated to the original domain score from 0.5 to 0.7*

If Rasch standardized residuals are applied to calculate the dimension coefficient, the cutting point at ≤0.55 can be the criterion for identifying the interlaced phenomenon in data (Chien, 2012).

We demonstrate how to simulate Rasch data with Winsteps and show that (1) the generation of responses under Rasch rating scale model (Andrich, 1978; Linacre, 2007); (2) the use of Winsteps to run the DOS commend: LResult = "START /wait ..\winsteps BATCH=YES cntrolefilel.txt outputfile.txt NI=" & Item2 & " CODES=" & mcode & " " & idelete & " " & rescore; (3) the execution of a batch file in MS Excel using the statement [Shell(mpath & "kidmap_bat.bat", 1)]; (4) the tables generated by Winsteps were used to read and write the required part extracted into the spreadsheet we need for next analyses.

Many Rasch learners hope to know the procedure and process of Rasch simulation with Winsteps. We provide a video at: https://youtu.be/0exg3mZVt6w to share our experience with RMT readers.

Tsair-Wei Chien[1]*, Yang Shao*[2]
Corresponding author: Tsair-Wei Chien
Email: smile@mail.chimei.org.tw
Affiliated institutes:
[1] Chi Mei Medical Center, Taiwan
[2] Tongji Zhejiang College, Jiaxing, China

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-73.

Chien, T.-W. (2012). Cronbach's alpha with the dimension coefficient to jointly assess a scale's quality. *Rasch Measurement Transactions*, 26(3), 1379.

Chien, T.-W, Shao, Y, & Jen, D.-H. (2017). Development of a Microsoft Excel tool for applying a factor retention criterion of a dimension coefficient to a survey on patient safety culture. *Health Qual Life Outcomes*, 15: 216.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Med Educ., 37*, 830-837.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed): *Educational Measurement*, 3rd Edition. New York, NY: American Council on Education and Macmillan.

Linacre, J. M. (2007). How to simulate Rasch data. *Rasch Measurement Transactions*, 21(3), 1125.

Linacre, J. M. (2010). Removing Rasch misfit and cleaning windows. *Rasch Measurement Transactions*, 23(4), 1241.

Tennant A., & Pallant J. F. (2006). Unidimensionality matters! (A tale of two Smiths?), *Rasch Measurement Transactions*, *20*(1), 1048-51.

Upcoming Rasch Measurement Courses and Workshops

August 14-16 2019, Workshop in Sweden with Richard Smith

On August 14-16, the national Swedish network for Psychometrics and Metrology in the health sciences (PMhealth; www.hkr.se/en/pmhealth) is organizing an interactive workshop led by Richard M. Smith under the theme *An Introduction to Rasch Measurement: Theory and Applications.*

As many of you are aware, Richard Smith has been working with the Rasch model for many years and has, among other things, made himself known as an expert on Rasch model fit, as the founding and current editor of the *Journal of Applied Measurement*, and as the organizer of the *International Outcome Measurement Conference* (IOMC).

The workshop will be held at Kristianstad University, Kristianstad, Sweden. Tentative workshop program together with practical information and registration is available at www.hkr.se/pmhealth2019rs. In addition to the workshop, there will also be room for informal discussions and networking.
Note that the number of participants is limited, on a first come first served basis.

July – December 2019, Online Course with David Andrich

The Course/Academic unit *Introduction to Classical and Rasch Measurement Theories* is being offered **on-line** again from July to December this year. The unit can be taken as professional development or towards a higher degree. The website for information is: http://www.education.uwa.edu.au/ppl/courses

August 24 – 31 and September 21 – 28, Introduction to Rasch Analysis in Mexico with Agustín Tristán (in Spanish)

Análisis de Rasch introductorio (en español) en Web. Agustín Tristán.
Instituto de Evaluación e Ingeniería Avanzada.
San Luis Potosí, México.
24 a 31 de agosto 2019 (24 to 31 August 2019). www.ieia.com.mx
21 a 28 de septiembre 2019 (21 to 28 September 2019). www.ieia.com.mx

November 23 – 30, Online Course with Agustín Tristán (in French)

Cours d'initiation à l'analyse de Rasch (en français) en Web. Agustin Tristan. Institut d'évaluation et d'ingénierie avancée. San Luis Potosi, Mexique.
23 au 30 novembre 2019. www.girief.org

List of Recent Publications in Journal of Applied Measurement

Vol. 20, No. 1, Spring 2019

The Effects of Probability Threshold Choice on an Adjustment for Guessing using the Rasch Model
Glenn Thomas Waterbury and Christine E. DeMars

Quantifying Item Invariance for the Selection of the Least Biased Assessment
W. Holmes Finch, Brian F. French, and Maria E. Hernandez Finch

Rasch Model Calibrations with SAS PROC IRT and WINSTEPS
Ki Cole

Student Perceptions of Grammar Instruction in Iranian Secondary Education: Evaluation of an Instrument using Rasch Measurement Theory
Stefanie A. Wind, Behzad Mansouri, and Parvaneh Yaghoubi Jami

Computer Adaptive Test Stopping Rules Applied to the Flexilevel Shoulder Functioning Test
Trenton J. Combs, Kyle W. English, Barbara G. Dodd, and Hyeon-Ah Kang

Examining Rater Judgements in Music Performance Assessment using Many-Facets Rasch Rating Scale Measurement Model
Pey Shin Ooi and George Engelhard, Jr.

Examining Differential Item Functioning in the Household Food Insecurity Scale: Does Participation in SNAP Affect Measurement Invariance?
Victoria T. Tanaka, George Engelhard, Jr., and Matthew P. Rabbitt

Accuracy and Utility of the AUDIT-C with Adolescent Girls and Young Women (AGYW) Who Engage in HIV Risk Behaviors in South Africa
Tracy Kline, Corina Owens, Courtney Peasant Bonner, Tara Carney, Felicia A. Browne, and Wendee M. Wechsberg

Vol. 20, No. 2, Summer 2019

Loevinger on Unidimensional Tests with Reference to Guttman, Rasch, and Wright
Mark H. Stone and A. Jackson Stenner

Standard-Setting Procedures for Counts Data
Rianne Janssen, Jorge González, and Ernesto San Martín

Expected Values for Category-To-Measure and Measure-To-Category Statistics: A Simulation Study
Eivind Kaspersen

Missing Data and the Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation
Glenn Thomas Waterbury
Cross-Cultural Comparisons of School Leadership using Rasch Measurement
Sijia Zhang and Stefanie A. Wind

Development of a Mathematics Self-Efficacy Scale: A Rasch Validation Study
Song Boon Khing and Tay Eng Guan

Lucky Guess? Applying Rasch Measurement Theory to Grade 5 South African Mathematics Achievement Data
Sarah Bansilal, Caroline Long, and Andrea Juan

A Note on the Relation between Item Difficulty and Discrimination Index
Xiaofeng Steven Li

Rasch Measurement SIG Business Meeting Minutes & SIG Announcements

The Rasch Measurement SIG Business Meeting was held on April 10, 2019 at the AERA Annual Conference in Toronto, Canada. There were 10 members in attendance. The topics of discussion included the website, SIG awards, and membership. With regards to the website, we discussed the possibility of canceling the Raschsig.org website and migrating it over to the AERA website because the AERA website would not cost the SIG any money. We also discussed the need to better publicize the call for nominations for the Rasch SIG awards and make clearer the requirements for each award. There was also a suggestion to publicize the speaker in advance of the business meeting in hopes of attracting more attendees. We brainstormed ideas to increase SIG membership which included reaching out to people who present at the Rasch SIG sessions and new students doing Rasch-related research. We ended the meeting with a presentation by Adrienne Walker, who was awarded the Georg William Rasch Measurement Early Career Publication Award at the 2017 AERA annual conference.

We look forward to next year's AERA annual conference in San Francisco, CA. Trent Haines and Courtney Donovan have volunteered to serve as the co-chairs for the 2020 meeting. Please consider submitting a proposal to the Rasch SIG. The deadline for submissions is July 10, 2019. Additionally we will be seeking nominations for officers and for the senior research award this fall.

Cari Hermann Abell