Impact of adjusting for inter-rater variability in conference abstract ranking and selection processes

Scanlan, Justin Newton; Lannin, Natasha A.; Hoffmann, Tammy; Stanley, Mandy; Mcdonald, Rachael

Published in:
Australian Occupational Therapy Journal

DOI:
10.1111/1440-1630.12440

TITLE:

Impact of adjusting for inter-rater variability in conference abstract ranking and selection processes

RUNNING HEAD:

Conference abstract ranking and selection

AUTHORS:

Justin Newton Scanlan

PhD, MHM, BOccThy

Senior Lecturer, The University of Sydney, Faculty of Health Sciences, Sydney NSW, Australia

Allied Health Research Support Leader, Sydney Local Health District, Mental Health Services, Sydney NSW, Australia

justin.scanlan@sydney.edu.au

Natasha A. Lannin

PhD, BSc(OT), GradDip

Associate Professor, La Trobe University, School of Allied Health (Occupational Therapy), Melbourne, Victoria, Australia

Associate Professor, Alfred Health (Occupational Therapy), Melbourne, Victoria, Australia

N.Lannin@latrobe.edu.au

Tammy Hoffmann

PhD, BOccThy (Hons 1)

Professor, Bond University, Centre for Research in Evidence-Based Practice, Faculty of Health Sciences and Medicine, Gold Coast, Queensland, Australia

thoffmann@bond.edu.au

Mandy Stanley

PhD, MHlthSc(OT), BHlthSc(OT)

Senior Lecturer, University of South Australia, Sansom Institute, School of Health Sciences, Adelaide, SA, Australia

mandy.stanley@unisa.edu.au

Rachael McDonald

PhD, GCHE, PGDip(Biomech), BAppSc(OT)

Associate Professor and Chair, Swinburne University of Technology, Department of Health and Medical Science

rachaelmcdonald@swin.edu.au

CORRESPONDENCE:

Justin Scanlan

Faculty of Health Sciences, C43J

The University of Sydney

PO Box 170

LIDCOMBE NSW Australia

Ph: 02 9351 9022

E: justin.scanlan@sydney.edu.au

Declaration of Authorship

NL conceived the original idea for the study. JNS, NL and TH devised and implemented the original study design for the project. TH, RM and MS supported and guided the implementation of the study reported in this paper. JNS completed the analyses presented in this paper and drafted the initial manuscript. All authors contributed to the interpretation of results and writing of the manuscript. The authors declare that there are no conflicts of interest.

Funding

This project received no specific funding.

ABSTRACT

Introduction: Scientific conferences provide a forum for clinicians, educators, students and researchers to share research findings. To be selected to present at a scientific conference, authors must submit a short abstract which is then rated on its scientific quality and professional merit and is accepted or rejected based on these ratings. Previous research has indicated that inter-rater variability can have a substantial impact on abstract selection decisions. For their 2015 conference, the Occupational Therapy Australia National Conference introduced a system to identify and adjust for inter-rater variability in the abstract ranking and selection process.

Method: Ratings for 1340 abstracts submitted for the 2015 and 2017 conferences were analysed using many-faceted Rasch analysis to identify and adjust for inter-rater variability. Analyses of the construct validity of the abstract rating instrument and rater consistency were completed. To quantify the influence of inter-rater variability of abstract selection decisions, comparisons were made between decisions made using Rasch-calibrated measure scores and decisions that would have been made based purely on raw average scores derived from the abstract ratings.

Results: Construct validity and the measurement properties of the abstract rating tool were good to excellent (item fit *MnSq* scores ranged from 0.8 to 1.2; item reliability index = 1.0). Most raters (24 of 27, 89%) were consistent in their use of the rating instrument. When comparing abstract allocations under the two conditions, 25% of abstracts (n = 341) would have been allocated differently if inter-rater variability was not accounted for.

Conclusion: This study demonstrates that, even with a strong abstract rating instrument and a small rater pool, inter-rater variability still exerts a substantial influence on abstract selection decisions. It is recommended that all occupational therapy conferences internationally, and

scientific conferences more generally, adopt systems to identify and adjust for the impact of inter-rater variability in abstract selection processes.

INTRODUCTION

Scientific conferences provide the opportunities for clinicians, educators, students and researchers to present papers and disseminate research findings. Conferences have a limited number of spaces available for podium presentations and the number of abstracts submitted typically far exceed the number of presentation spaces available. Therefore, it is imperative that conference scientific program committees implement fair and transparent systems for judging the relative quality and merit of each abstract and determining which are selected for presentation.

Most conferences use a peer review approach to abstract ratings. In this process, a selection of peer reviewers are appointed and each asked to rate a number of abstracts. These ratings may be guided by a structured abstract rating tool or based on global, overall ratings of quality and scientific merit. The process of abstract rating and ranking has been the topic of considerable research (Bhandari, Templeman, & Tornetta, 2004; Cohen & Patel, 2006; Montgomery, Graham, Evans, & Fahey, 2002; Poolman et al., 2007; Rowe et al., 2006; Scanlan, Lannin, & Hoffmann, 2015; Timmer, Sutherland, & Hilsden, 2003; van Mastrigt & Downie, 1994). Issues have been identified in terms of poor inter-rater agreement on abstract ratings (Bhandari et al., 2004; Montgomery et al., 2002; Rowe et al., 2006), the use of large rater pools (Bhandari et al., 2004; van Mastrigt & Downie, 1994) and the use of unstructured, "global" approaches to abstract ratings (Montgomery et al., 2002; Poolman et al., 2007; Rowe et al., 2006; van Mastrigt & Downie, 1994).

The abstract rating and ranking processes used in occupational therapy conferences has sparked some debate in the professional literature, especially regarding the use of large pools

of raters, variability between raters and the rejection of abstracts reporting on high quality research studies (Hammell, 2009; Lannin et al., 2009; Pattison & Pascoe, 2009). Following this, improvements were made to the abstract rating and ranking process for the 2011 Occupational Therapy Australia National Conference, including: (i) the use of a structured abstract rating tool; (ii) the use of a limited rater pool to rate all abstracts submitted; and (iii) abstract rankings based on average scores generated from ratings completed by a large number of raters (Lannin & Scanlan, 2013).

Although this new system addressed many of the concerns raised in the debate (Lannin et al., 2009; Pattison & Pascoe, 2009), rater burden was high (with each rater rating approximately 150 abstracts) and variability between raters continued to have a small, but substantial impact on abstract disposition (i.e., being selected for a long oral [12 minutes], short oral [5 minutes] or poster presentation) (Scanlan et al., 2015). Using many-faceted Rasch analysis (Linacre, 2014) to model the impact of inter-rater variability on abstract disposition, the authors identified that approximately 20% of abstracts would have been selected for different presentation types if rater variability had been taken into account (Scanlan et al., 2015). This suggested that, even with these improved systems, some abstracts were advantaged by being allocated to raters who tended to give high ratings and some were disadvantaged by being allocated to raters who tended to give lower ratings.

Rater variability of the type described above is common in “expert judgment” situations where raters tend to use “internalised criteria” to guide scoring (Myford & Wolfe, 2004). While many attempts have been made to improve inter-rater consistency through training and other approaches, these attempts tend to be unsuccessful and may inadvertently reduce raters’ overall consistency (Lunz & Stahl, 1993b; Lunz, Stahl, & Wright, 1994).

Commencing with the 2015 Occupational Therapy Australia national conference, the many-faceted Rasch analysis procedure used in the modelling study (Scanlan et al., 2015) was implemented during the abstract ranking process to adjust for inter-rater variability and guide abstract selection. This study aimed to investigate the validity of the overall abstract ranking process used and to quantify the impact that adjusting for inter-rater variability had on overall abstract selection. The key research questions were: (1) does the abstract rating instrument used demonstrate good construct and internal validity?; (2) do individual raters use the abstract rating instrument consistenty (intra-rater reliability / consistency)?; and (3) what influence does inter-rater variability exert on overall abstract selection decisions?

METHOD

This study was approved by the University of Sydney Human Research Ethics Committee (approval numbers 2014/1026 for the 2015 data and 2016/774 for the 2017 data). Data consisted of abstract ratings completed by members of the 2015 and 2017 Conference Scientific Program Committees. Data were stripped of all identifying information in relation to individual raters prior to being provided to the research team. As the data were non-identifiable and the project was considered by the Human Research Ethics Committee as a secondary analysis of existing data, individual informed consent was not required. However, all members of the committees were aware that their de-identified abstract ratings would be analysed using the many-faceted Rasch analysis approach and gave their informal consent for this to occur through the chair of the Conference Scientific Program Committee.

Abstract rating process

Individuals wishing to make presentations at the conference were required to submit structured abstracts of a maximum of 250 words. A total of 662 abstracts were submitted for the 2015 conference and 678 for the 2017 conference. All abstracts were rated by members of the conference scientific program committee. This committee had 13 members for the 2015 conference and 14 members for the 2017 conference. Four members served on both the 2015 and 2017 committees. Committee members were selected by merit based on expressions of interest addressing six essential and two desirable criteria. Essential criteria were: (i) Occupational Therapy Australia Membership; (ii) High level of research experience (PhD or equivalent); (iii) Peer reviewed publications/s in nominated area of expertise; (iv) Demonstrated ability to work independently and perform effectively as part of a team; (v) Demonstrated high standard of oral and written communication skills, time management, and organisational skills to ensure effective teamwork; and (vi) Able to attend monthly teleconferences and possibly one face to face meeting. Desirable criteria were: (i) Previous experience in, and current knowledge of, contributing to the planning and delivery of a conference or other professional development event; and (ii) Experience / insight / awareness related to current and future topics or issues important for occupational therapy practice, education and research.

For the 2015 conference, each abstract was rated by three members of the committee and for the 2017 conference, this was reduced to two members, based on recommendations from the modelling study to minimise rater burden (Scanlan et al., 2015). In both 2015 and 2017 conferences a small proportion of abstracts were rated by all members of the committee to create better cross-connections within the data set.

A structured abstract rating instrument was used to rate all abstracts. This instrument was originally developed by the Canadian Association of Occupational Therapists for use in their national conferences. It was acquired and adopted as part of the improvements made to the Occupational Therapy Australia conference processes in 2011 (Lannin & Scanlan, 2013). This instrument included 10 items in three categories (Quality of the presentation content [5 items]; Educational value [3 items]; and Quality of the written abstract [2 items]). Items in the first section were rated on a 5-point scale (1 = unacceptable; 2 = marginal; 3 = acceptable; 4 = good; 5 = exceptional) and items in the second and third sections were rated on a 5-point scale with anchors at 1 (low, unacceptable); 3 (moderate, acceptable); and 5 (high, exceptional). Previous research has demonstrated that this structured rating instrument has sound measurement properties, including rating scale functioning, construct validity and internal reliability (Scanlan et al., 2015). No changes were made to the instrument between the 2015 and 2017 conferences. A copy of the structured abstract rating instrument is available as an online only appendix to this paper ***[insert link to online only appendix].***

The overall aim of the abstract rating process was to assess the quality of each abstract to determine a ranking to be used to select abstracts for the various presentation formats. For the 2015 conference, there were 180 presentation slots for long papers and 180 slots for short papers. For the 2017 conference, this was revised to 193 long paper slots and 131 short paper slots. Abstracts of sufficient scientific quality that were not allocated to long or short paper presentation slots were offered poster presentations. Authors could also indicate a preference for a poster presentation and in this case, the abstract was allocated a poster presentation regardless of overall ranking.

Analyses

A number of analyses were completed based on the various research questions guiding the study. The procedures used for these analyses are described below. The main program used for analyses presented in this paper is the many-faceted Rasch analysis program, FACETS (Version 3.70.1: Linacre, 2014). Many-faceted Rasch analysis is described in detail elsewhere (Bond & Fox, 2015; Linacre, 2014), however, in short, this approach to analysis allows for the simultaneous consideration of several factors that influence measurement. In the context of this study, these elements were the severity of the rater, and the scores allocated to the abstract using the abstract rating instrument.

Internal and construct validity of the abstract rating instrument

Using guidelines set out in existing literature (W. P. Fisher, 2007; Linacre, 1999), we evaluated several aspects of internal and construct validity of the abstract rating instrument. The first step was to investigate item functioning; for an item to be useful, it must be related to the construct being measured (in this case, scientific quality of the abstract) and must be used consistently. We report fit statistics to give an indication of how consistently the item is being used. Low fit statistics indicate that an item might not provide useful information for determining the scientific quality of an abstract while high fit statistics suggest that the item is unclear or is being used in unpredictable ways (Bond & Fox, 2015).

Fit statistics will be presented as Mean Square (*MnSq*) scores. General guidelines suggest *MnSq* scores in the range of 0.8 and 1.3 are “excellent” (W. P. Fisher, 2007) and scores in the range of 0.5 and 1.5 are “acceptable” (Linacre, 1999). Fit statistics outside of these ranges will be used to suggest that the item may not measure scientific quality and, in the case of very high fit statistics, that the item’s inclusion may degrade the overall measurement model. Secondly, point-measure correlations will be calculated to ensure the item contributes to the

overall construct. Negative point-measure correlations of an item will be highlighted to suggest that the item is not part of the construct being measured (Bond & Fox, 2015).

Following this, rating scale functioning was investigated. If each rating scale in the abstract rating instrument was functioning optimally, then each point on the scale would represent a unique and meaningful aspect of scientific quality (i.e., the construct being measured) (Bond & Fox, 2015). Firstly, average measure scores should progress monotonically (that is, abstracts being rated a four on any item should, on average, have higher measure scores for overall scientific quality than abstracts receiving ratings of three, and so on). Secondly, Rasch-Andrich thresholds should also progress monotonically and, optimally, should have step progressions of between 1.4 to 5.0 logits (Linacre, 1999). Where Rasch-Andrich thresholds do not progress monotonically or do not demonstrate sufficient step progression, this suggests that the categories may not represent a unique or distinguishable progression of person abilities. In this case, it will be used to suggest that the scale contains too many categories and the rating scale categories may need to be collapsed (Linacre, 1999).

Finally, Rasch-generated item separation statistics and reliability indices will be calculated to provide an overall indication of the construct validity of the abstract rating instrument. The separation statistic provides an indication of the ability of the instrument to separate people (or in this case, abstracts) into statistically distinguishable groups based on the construct under investigation. Separation statistics of > 5.0 are considered “excellent” (W. P. Fisher, 2007). The reliability index will also be reported as a measure of internal consistency (conceptually equivalent to Cronbach’s α). Reliability indices of > .91 will be considered “excellent” (W. P. Fisher, 2007).

Rater consistency in using the abstract rating instrument

Before exploring the influence of variations between raters, it was important to ensure that all raters were using the instrument consistently across the various different abstracts they were rating. To investigate this, rater fit statistics were examined. Acceptable ranges of fit statistics for raters have not been well-established, although it is generally considered that the range of 0.6 to 1.5 is acceptable (Lunz & Stahl, 1993a; Lunz, Wright, & Linacre, 1990). Rater fit *MnSq* scores of > 1.5 suggest that there is some degree of unpredictability in their use of the rating scale and *MnSq* scores of > 2.0 may degrade the overall measurement model (Bond & Fox, 2015).

Influence of inter-rater variability on abstract allocation

The degree of inter-rater variability can be evaluated through examination of the separation statistic and reliability index generated by the many-faceted Rasch analysis. If the influence of inter-rater variability is low, then the separation statistic would be low (< 2.0) and the reliability index would also be low (< .67) (W. P. Fisher, 2007). If the influence of inter-rater variability is more substantial, then the separation statistic would be high (> 5.0) as would the reliability index (> .94) (W. P. Fisher, 2007).

To examine the "real-world" influence of inter-rater variability in this context, abstract selection decisions (allocation to a long paper, short paper or poster) were examined under two different conditions. The first condition was to rank abstracts using mean scores for each abstract based on scores allocated by the various raters. The second condition was to rank abstracts according to the person measure scores derived from the many-faceted Rasch analysis, which corrected for inter-rater variability. Under both conditions, abstracts were allocated to presentation types by ranking according to the number of presentation slots at

each conference. Comparing abstract allocations under these two conditions allowed for the identification of the number of abstracts that would have been selected for a different presentation type had inter-rater variability not been taken into account.

As an award is presented for the top-rated abstract, this was also investigated to determine if accounting for inter-rater variability changed this ranking. Other changes such as change to abstracts ranked in the top 10 and top 50 of all abstracts and absolute change in rank were also investigated, although these have negligible impact on the overall outcome of the abstract selection process.

RESULTS

Results from the analysis of item functioning are presented in Table 1. Fit statistics for all items over both data sets met the criteria set for this study and all fell within the range of 0.8 to 1.2. Point-measure correlations were all positive and $\geq .5$. The item separation statistics for 2015 and 2017 were 12.9 and 8.6 respectively and reliabilities indices were 1.0. This suggests that, overall, all items are used consistently, contribute to the overall construct of scientific quality and contribute to separating abstracts into a large number of statistically distinguishable groups.

Rating scale functioning data is presented in Table 2. Average measure scores and Rasch-Andrich thresholds all progressed monotonically. Step progressions between thresholds ranged from 1.2 to 2.5 logits. These data suggest that, overall, the rating scales are operating effectively and that each category on each scale represented a distinguishable range of performance on the construct under examination.

Quality control statistics for raters are presented at the top of Table 3. Individual rater fit statistics ranged from 0.7 to 1.6, with 89% of raters' fit statistics falling within the range of 0.6 to 1.4. These results suggest that, in the vast majority of cases raters were consistent in their use of the abstract rating instrument.

The separation statistics (13.3 for 2015 and 17.4 for 2017) and reliability indices (1.0 for both 2015 and 2017) for raters were very high (Table 3). This suggests that, despite each rater being consistent, there was substantial variation between individual raters themselves.

Data on changes in abstract allocation and other differences in ranking between rankings based on mean scores and Rasch-calibrated measure scores are presented in Table 4. Allocation outcomes based on Rasch-calibrated measure scores compared with those based on mean scores were different for 18% of abstracts for the 2015 conference and for 33% of abstracts for the 2017 conference. In both cases, the top ranked abstract using Rasch-calibrated measures scores was different to the top ranked abstract using mean scores. There were also substantial proportions of abstracts that had changes in rank of greater than 50 or 100 places when comparing the two ranking systems.

DISCUSSION

This study was established primarily to investigate the influence of inter-rater variability on the decision-making processes underpinning abstract selection in the Occupational Therapy Australia National Conference. Additionally, the study investigated the construct validity (item functioning and rating scale functioning) of the structured abstract rating instrument and the consistency of individual raters. Overall, results from this study demonstrate that the

measurement properties of the abstract rating instrument are good to excellent and that, in general, raters are consistent in their use of the instrument. However, despite these positive factors, variability between raters could still have a substantial impact on overall abstract rankings.

These results have implications for abstract ranking and rating processes for all scientific conferences, within occupational therapy and across all scientific disciplines. Although inter-rater variability in abstract ratings has been identified numerous times in the scientific literature (Bhandari et al., 2004; Cohen & Patel, 2006; Montgomery et al., 2002; Rowe et al., 2006), it is an issue that is generally overlooked in the abstract ranking process for most scientific conferences. Results from this study suggest that overlooking this issue could undermine the fairness, equity and quality of scientific conferences. Without considering and correcting for the influence of between rater variability, there is risk that some abstracts with lower merit will be elevated to more influential presentation types and abstracts with higher merit may not receive the profile they deserve or be rejected.

In discussing the issue of inter-rater variability, it must be reiterated that this is not due to "unfair" or "biased" rating practices on the part of the raters. Data from this study and the previous modelling study (Scanlan et al., 2015) demonstrate that the vast majority of raters were very consistent (and therefore "fair") within their ratings. The issue emerges from variations between individual raters. What one rater considers "exceptional," another might consider "good" or even only "acceptable." In other words, some raters may be more "severe" while others may be more "lenient." In expert rating situations, this is referred to as each rater's "internalised criteria." Whereas in some measurement situations (e.g., measuring height or limb circumference), rater skill developed through training can enhance inter-rater

consistency, in rating situations requiring expert judgements, rater training is far less successful, consistent and precise (Barrett, 2001; Lumley & McNamara, 1995). Expert raters have generally developed their expertise over a large number of years and therefore have strong "internalised criteria" which tend to be difficult to alter (Lumley & McNamara, 1995; Lunz & Stahl, 1990). Indeed, attempts to alter rater scoring in expert judgement situations may actually have the unintended consequence of making raters less consistent. In this situation, raters may "second guess" their own internalised criteria and may oscillate between rating according to their internalised criteria and how they believe they are supposed to be rating.

Given the factors described above, many authors (Linacre, 1989; Lumley & McNamara, 1995; Lunz et al., 1990; Montgomery et al., 2002) have suggested that the best way to manage inter-rater variability in expert judgement contexts is to accept that it will be present and then identify and account for it. This has been applied in educational examination contexts as well as in an assessment familiar to many occupational therapists, the Assessment of Motor and Process Skills (AMPS) (A. G. Fisher, 1993; Tesio, 2003).

In considering the overall results of this study, some additional aspects are also worthy of discussion. These include the small number of raters whose fit statistics fell slightly outside of the generally acceptable range and the substantial differences in abstract ranking changes seen between the 2015 and 2017 data sets. In terms of raters whose fit statistics fell outside of the generally acceptable range of 0.6 to 1.5, there were two in 2015 and one in 2017. In all cases, the fit statistics for these raters were close to the generally accepted range (all fit statistics were ≤ 1.6) and were not in the range likely to denigrate the overall measurement

model. However, the slightly inconsistent ratings provided by these raters may have influenced the ratings of individual abstracts.

To provide an additional check, those abstracts rated by raters whose fit statistics were > 1.5 that were located around the "critical cut points" (i.e., the thresholds between "long paper" and "short paper" and "short paper" and "poster") were further examined by other members of the scientific program committee to ensure the eventual allocation to presentation type appeared appropriate. Although specific records were not maintained, this additional evaluation resulted in very few changes to the overall allocation.

Finally, when comparing the data presented in Table 4 for 2015 and 2017 data, it is clear that the influence of inter-rater variability was more substantial in 2017. The reason for this is likely due to a change made where for the 2015 conference, each abstract was rated by three members of the scientific program committee, whereas for the 2017 conference each abstract was rated by only two members. As was noted in the modelling study (Scanlan et al., 2015), the influence of inter-rater variability is more substantial in rating situations involving fewer raters per abstract.

This is simply due to probability. In the three rater situation of 2015, the probability of an abstract being allocated to three of the five "hardest" (or "easiest") raters was 3.5%, whereas in the two rater situation of 2017, the probability of an abstract being allocated to two of the five "hardest" (or "easiest") raters was 11.0%. This means that it should be expected that two to three times more abstracts would be influenced by inter-rater variability in 2017 when compared with 2015. This does not suggest that the variations between raters were more pronounced in 2017 (although there is some evidence in terms of rater separation statistics

that this may have been the case), but moreso a reflection of the increased probability of an abstract being allocated to raters who are all amongst the “hardest” or “easiest” raters in the rater pool. While it is true that increasing the number of raters rating each abstract assists to reduce the impact of inter-rater variability, even in the case of having five raters rating each abstract, inter-rater variability still exerted a substantial influence on abstract raking (Scanlan et al., 2015). Additionally, systems of having large numbers of raters rating each abstract creates additional rater burden which can introduce greater inconsistencies in ratings due to rater fatigue (Wolfe, Moulder, & Myford, 2001).

Conclusion

This study has highlighted the importance of considering and correcting for inter-rater variability in the context of abstract ranking processes in scientific conferences by quantifying the number of abstract allocations decisions influenced by inter-rater variability. This has important implications for all scientific conferences. Given the results from this study, it is recommended that a similar approach to the abstract ranking process be adopted for all occupational therapy conferences internationally and, more broadly, all scientific conferences to address issues of fairness and transparency.

Key Points for Occupational Therapy

- Even with a strong abstract rating instrument and a small number of raters, abstract ranking processes are substantially influenced by variability between raters
- Occupational therapy scientific conferences should adopt systems to identify and adjust for inter-rater variability in the abstract ranking process
- Many-faceted Rasch analysis is one way that inter-rater variability can be accounted for.

REFERENCES

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*, 49-58.

Bhandari, M., Templeman, D., & Tornetta, P. (2004). Interrater reliability in grading abstracts for the Orthopaedic Trauma Association. *Clinical Orthopaedics and Related Research, 423*, 217-221. doi:10.1097/01.blo.0000127584.02606.00

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Cohen, I. T., & Patel, K. (2006). Peer review interrater concordance of scientific abstracts: A study of anesthesiology subspecialty and component societies. *Anesthesia & Analgesia, 102*, 1501-1503. doi:10.1213/01.ane.0000200314.73035.4d

Fisher, A. G. (1993). The assessment of IADL motor skills: An application of many-faceted Rasch analysis *American Journal of Occupational Therapy, 47*, 319-329. doi:10.5014/ajot.47.4.319

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*, 1095.

Hammell, K. W. (2009). Further concerns for occupational therapy conferences. *Australian Occupational Therapy Journal, 56*, 216. doi:10.1111/j.1440-1630.2009.00805.x

Lannin, N. A., Gustafsson, L., Cusick, A., Walker, M., Steultjens, E., Fricke, J., . . . Wallen, M. (2009). Scholarly communication and concerns for our conferences. *Australian Occupational Therapy Journal, 56*, 147-148. doi:10.1111/j.1440-1630.2009.00786.x

Lannin, N. A., & Scanlan, J. N. (2013). Enhancing the quality of scoring of abstracts submitted to an occupational therapy conference. *Australian Occupational Therapy Journal, 60*(s1), 101. doi:10.1111/1440-1630.12062

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103-122.

Linacre, J. M. (2014). *A User's Guide to FACETS Rasch-Model Computer Programs. Program Manual 3.71.4*: Available from www.winsteps.com.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*, 54-71. doi:10.1177/026553229501200104

Lunz, M., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions, 13*, 425-444. doi:10.1177/016327879001300405

Lunz, M., & Stahl, J. A. (1993a). The effect of rater severity on person ability measure: a Rasch model analysis. *American Journal of Occupational Therapy, 47*, 311-317. doi:10.5014/ajot.47.4.311

Lunz, M., & Stahl, J. A. (1993b). Impact of examiners on candidate scores: An introduction to the use of multifacet Rasch model analysis for oral examinations. *Teaching and Learning in Medicine: An International Journal, 5*, 174-181. doi:10.1080/10401339309539614

Lunz, M., Stahl, J. A., & Wright, B. D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54*, 913-925. doi:10.1177/0013164494054004007

Lunz, M., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345. doi:10.1207/s15324818ame0304_3

Montgomery, A. A., Graham, A., Evans, P. H., & Fahey, T. (2002). Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference. *BMC Health Services Research, 2*(8). doi:10.1186/1472-6963-2-8

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 460-517). Maple Grove, MN: JAM Press.

Pattison, M., & Pascoe, J. (2009). Response to 'Scholarly communication and concerns for our conferences'. *Australian Occupational Therapy Journal, 56,* 214-215. doi:10.1111/j.1440-1630.2009.00797.x

Poolman, R. W., Keijser, L. C. M., Malefijt, A. C. d. W., Blankevoort, L., Farrokhyar, F., & Bhandari, M. (2007). Reviewer agreement in scoring 419 abstracts for scientific orthopedics meetings. *Acta Orthopaedica, 78,* 278-284. doi:10.1080/17453670710013807

Rowe, B. H., Strome, T. L., Spooner, m. C., Blitz, S., Grafstein, E., & Worster, A. (2006). Reviewer agreement trends from four years of electronic submission of conference abstract. *BMC Medical Research Methodology, 6*(14). doi:10.1186/1471-2288-6-14

Scanlan, J. N., Lannin, N. A., & Hoffmann, T. (2015). Can Rasch analysis enhance the abstract ranking process in scientific conferences? Issues of interrater variability and abstract rating burden. *Journal of Continuing Education in the Health Professions, 35,* 18-26. doi:10.1002/chp.21263

Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine, 35,* 105-115. doi:10.1080/16501970310010448

Timmer, A., Sutherland, L. R., & Hilsden, R. J. (2003). Development and evaluation of a quality score for abstracts. *BMC Medical Research Methodology, 3*(2). doi:10.1186/1471-2288-3-2

van Mastrigt, R., & Downie, J. W. (1994). Statistical evaluation of the function of the 1992 International Continence Society Scientific Committee. *Neurology and Urodynamics, 13,* 323-331. doi:10.1002/nau.1930130314

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2,* 256-280.

Table 1. Item statistics, presented in measure order for 2017 data

	2015 data					2017 data				
Item[†]	Measure	Model SE	Infit *MnSq*	Outfit *MnSq*	Pt-Meas Corr	Measure	Model SE	Infit *MnSq*	Outfit *MnSq*	Pt-Meas Corr
1d	0.71	0.03	1.09	1.11	.68	0.55	0.03	0.98	0.97	.68
1e	0.66	0.03	0.99	1.02	.68	0.36	0.03	0.98	0.98	.70
1c	0.40	0.03	0.93	0.96	.66	0.30	0.03	1.01	1.02	.68
3	-0.25	0.03	0.96	0.97	.67	0.04	0.04	1.05	1.04	.66
4	-0.27	0.03	1.05	1.05	.63	-0.02	0.04	1.09	1.08	.64
5	-0.03	0.03	0.82	0.82	.71	-0.04	0.04	0.80	0.80	.74
1b	0.08	0.03	1.10	1.12	.60	-0.10	0.03	1.09	1.10	.65
2	-0.54	0.03	0.98	0.98	.61	-0.22	0.04	0.99	0.99	.61
6	-0.42	0.04	0.87	0.88	.70	-0.40	0.04	0.89	0.88	.70
1a	-0.36	0.03	1.19	1.20	.54	-0.48	0.04	1.12	1.14	.62
	Item separation: 12.87 Item strata: 17.50 Item reliability index: .99					Item separation: 8.60 Item strata: 11.80 Item reliability index: .99				

[†] Items: 1a) Introduction OR Rationale; 1b) Objectives (of project and/or presentation); 1c) Methods OR Approach; 1d) Results OR Practice implications; 1e) Conclusions; 2) Interest and appeal to audience; 3) Important contribution to practice, research, theory or knowledge; 4) Novel or innovative contribution (e.g., current trends or new ideas); 5) Abstract is self-contained; 6) Abstract is coherent and readable.
SE = Standard Error; *MnSq* = Mean Square; Pt-Meas Corr = Point-Measure Correlation.

Table 2. Rating scale functioning statistics

	2015 Data					2017 Data				
Rating	Avg Meas	Exp Meas	Outfit *MnSq*	R-A Meas	R-A SE	Avg Meas	Exp Meas	Outfit *MnSq*	R-A Meas	R-A SE
Section 1: Quality of Presentation Content										
1 unacceptable	-1.61	-1.99	1.6			-1.19	-1.49	1.5		
2 marginal	-0.75	-0.68	1.0	-2.38	0.06	-0.54	-0.45	0.9	-2.06	0.07
3 acceptable	0.20	0.21	1.1	-1.14	0.03	0.36	0.37	1.0	-0.86	0.04
4 good	1.02	1.04	1.0	0.59	0.02	1.19	1.18	1.0	0.62	0.03
5 exceptional	1.87	1.87	1.0	2.93	0.04	1.98	2.00	1.0	2.31	0.04
Section 2: Educational Value										
1 low	-2.08	-2.12	1.0			-1.69	-1.57	0.9		
2 ⇳	-0.47	-0.34	0.8	-3.14	0.13	-0.39	-0.36	0.9	-2.49	0.12
3 moderate	0.73	0.70	1.1	-1.10	0.05	0.54	0.51	1.0	-1.04	0.05
4 ⇳	1.55	1.53	1.0	1.00	0.03	1.39	1.35	1.0	0.68	0.04
5 high	2.22	2.31	1.1	3.24	0.05	1.97	2.15	1.1	2.85	0.05
Section 3: Quality of Written Abstract										
1 low	-2.52	-2.13	0.7			-1.85	-1.57	0.8		
2 ⇳	-0.54	-0.46	0.9	-2.80	0.15	-0.34	-0.29	0.9	-2.64	0.16
3 moderate	0.49	0.56	0.8	-1.36	0.06	0.47	0.60	0.8	-0.98	0.07
4 ⇳	1.46	1.41	0.9	0.83	0.04	1.53	1.46	0.8	0.68	0.05
5 high	2.38	2.22	0.9	3.33	0.06	2.38	2.28	0.9	2.93	0.06

Avg Meas = Average Measure; Exp Meas = Expected Measure; *MnSq* = Mean Square; R-A Meas = Rasch-Andrich Threshold Measure; R-A SE = Rasch-Andrich Threshold Standard Error.

Table 3. Summary of rater statistics

Rater quality criterion	2015 Data	2017 Data
Number of raters	13	14
Individual rater consistency measures		
Rater infit *MnSq* range	0.73 to 1.56	0.72 to 1.60
Rater outfit *MnSq* range	0.73 to 1.53	0.71 to 1.53
Raters in infit *MnSq* range 0.6 to 1.5	11 (84.6%)	13 (92.8%)
Raters in outfit *MnSq* range 0.6 to 1.5	11 (84.6%)	13 (92.8%)
Raters in infit *MnSq* range 0.8 to 1.2	8 (61.5%)	10 (71.4%)
Raters in outfit *MnSq* range 0.8 to 1.2	8 (61.5%)	10 (71.4%)
Inter-rater variability measures		
Rater separation index	13.25	17.36
Rater strata	18.00	23.47
Rater reliability index	.99	1.00
Exact agreement	39.1%	32.6%
Model expected exact agreement	34.4%	33.1%
Rater measure scores range	-1.08 to 0.92	-1.18 to 1.62
Person measure score range	-5.11 to 3.55	-6.02 to 3.43
Rater measure range as a proportion of person measure range	23.1%	29.6%

Table 4. Changes in rank, average score to Rasch-modelled measure score

	2015 Data (n = 662)	2017 Data (n = 678)
Changes in presentation format[†]		
Poorer ranking		
Long paper to poster	0 (0.0%)	20 (2.9%)
Long paper to short paper	37 (5.6%)	33 (4.9%)
Short paper to poster	27 (4.1%)	63 (9.3%)
Better ranking		
Poster to long paper	0 (0.0%)	9 (1.3%)
Short paper to long paper	28 (4.2%)	42 (6.2%)
Poster to short paper	26 (3.9%)	56 (8.3%)
Changes in top abstracts		
Change in top-ranked abstract	Yes	Yes
Poorer ranking		
No longer in Top 10 abstracts	4 (0.6%)	8 (1.2%)
No longer in Top 50 abstracts	15 (2.3%)	23 (3.4%)
Better ranking		
Now in Top 10 abstracts	4 (0.6%)	8 (1.2%)
Now in Top 50 abstracts	17 (2.6%)	21 (3.1%)
Absolute change in rank		
Poorer ranking		
Change in rank >100 lower	68 (10.3%)	150 (22.1%)
Change in rank > 50 lower	125 (18.9%)	215 (31.7%)
Change in rank > 20 lower	211 (31.9%)	283 (41.7%)
Better ranking		
Change in rank >100 higher	27 (4.1%)	134 (19.8%)
Change in rank >50 higher	121 (18.3%)	199 (29.4%)
Change in rank >20 higher	224 (33.8%)	243 (35.5%)
Combined totals		
Change in rank >100 (total)	95 (14.4%)	284 (41.9%)
Change in rank >50 (total)	246 (37.2%)	414 (61.1%)
Change in rank >20 (total)	435 (65.7%)	526 (77.6%)

[†] For 2017 data, there were 193 long paper (12 minute oral presentation) slots and 131 short paper (5 minute oral presentations) slots; For 2015 data there were 180 long paper slots and 180 short paper slots.

[‡] This refers to change in the single abstract ranked highest. This is important as the top-ranked paper is awarded a prize.