Rasch Measurement and Patient Reported Value Claims: A Primer for Hemophilia

Paul C Langley, PhD, Adjunct Professor
College of Pharmacy, University of Minnesota, Minneapolis, MN

ABSTRACT

This commentary proposes that Rasch Measurement Theory (RMT) is an innovative method for assessments of patient-centric therapy response in hemophilia A and B, as they are in other disease states or target patient populations. RMT is a necessary and sufficient approach to moving from ordinal observations to interval measurement, which has arithmetic properties. This applies across the board in hemophilia and other disease states for clinical value claims, patient centric or subjective value claims as well as those for anticipated drug utilization and other medical care resources. The purpose of this commentary is to point out limitations regarding current methods for making claims regarding hemophilia response and to propose a new start in hemophilia studies to identify core claims that meet required measurement standards. This applies to both the development of new patient reported outcome instruments as well as the evaluation of existing instruments, with a focus on polytomous instruments and their sub-domains, to evaluate their possible application as measures that approximate RMT requirements.

Keywords: hemophilia, Rasch measurement, false outcome claims, ICHOM

INTRODUCTION

Current methods for evaluating evidence regarding gene therapies in Hemophilia A and B raises a number of critical questions regarding methodological limitations in applying assumption driven models for non-evaluable cost effectiveness claims [1]. Critical evaluation regarding assumption driven cost-per-quality adjusted life year (QALY) simulations, as the centerpiece of the current standard in health technology assessment (HTA), have been published [2 3 4]. The case against modeled claims is that they fail the standards of normal science and, in particular, the standards for fundamental measurement. The purpose of this brief commentary is to make clear that the issue of the standards for fundamental measurement, Rasch Measurement Theory (RMT) [5 6], has been ignored, not only in hemophilia studies, but more widely for patient relevant or centric outcome measures in hemophilia [7]. Three questions are raised: (i) is there any merit to the application of generic multiattribute scores in evaluating respondent status and response to therapy in hemophilia in terms of quality adjusted life years (QALYs); (ii) is it possible to ignore standards for fundamental measurement in proposing patient relevant health outcomes in hemophilia; and (iii) can we set fundamental measurement standards for patient reported outcome measures in hemophilia that capture more adequately the needs of patients and caregivers in hemophilia? It is important to recognize that RMT is not an option; it is the only valid basis for establishing patient reported outcome (PRO) claims. This has been recognized in health technology assessment since the early 1990s for disease specific claims and since the late 1970s in the development of the Nottingham Health Profile [8]. The key reference is the 1989 paper by Wright and Linacre where the make clear the importance of distinguishing between observations which are always ordinal and measurements which must be interval [9]. Measurement must be deduced from observations by applying rules. Rasch Measurement Theory (RMT) provides the tools for both dichotomous and polytomous instruments and the observations reported.

Indeed, RMT is the only model or set of rules that provide the necessary and sufficient transformations, if possible, between observations and interval measures of patient centric or subjective value claims. First proposed in 1953, RMT provides the solution for constructing PRO measures which retain their quantitative or calibration status irrespective of their application, while recognizing that that the measure must accommodate the interaction between the object to be measured and the measuring instrument [10]. The inherent unpredictability of this interaction led to a probabilistic interpretation when an individual responds to an item in a questionnaire; the probability of positively responding to an item is a mathematical function of the difference between the item's relative difficulty and the ability of the respondent to realize that difficulty for dichotomous response; where the response is polytomous, the Rasch model establishes the relative difficulty of each item stem recording the development of difficulty within that item as the rating scale has a number of thresholds and we need to model the likelihood of failure and success within each threshold [5 6].

VALUE CLAIMS STANDARDS

There are three standards that value claims must meet if they are to be accepted in hemophilia and other disease states by a formulary committee or health system [2]. These are:

- All value claims must refer to single attributes that meet the demarcation standards for normal science: they must be credible, evaluable and replicable

Corresponding author: Paul C. Langley, PhD
Adjunct Professor, College of Pharmacy
University of Minnesota, Minneapolis, MN
Director, Maimon Research LLC; Tucson, AZ
Email: langley@maimonresearch.com
Website: www.maimonresearch.com

DOI: https://doi.org/10.24926/iip.v13i4.5111

- All value claims must be consistent with the limitations imposed by the axioms of fundamental measurement: they must meet interval or ratio standards
- All value claims must be submitted with a protocol detailing how they are to be empirically evaluated in a meaningful timeframe

Failure to meet these standards must result in a claim being rejected. Accepting these standards means the end of assumption driven imaginary model simulations and attempts to pass off non-evaluable claims for cost-effectiveness as meaningful decision tools in hemophilia. Insisting on these standards means all claims are consistent with the credibility, evaluability and replication standards of normal science, including fundamental measurement, supporting hypothesis testing and progress in our understanding of the merits of hemophilia interventions.

Most importantly for consistency, PRO or subjective value claims in hemophilia, whether expressed in dichotomous or polytomous form, must meet the standards of RMT. This is the only framework for creating interval scales to evaluate response to therapy from questionnaires with either dichotomous or polytomous item responses. RMT applies in two contexts: (i) instrument development and (ii) the evaluation of existing instruments. In both cases the object is to create a metric that approximates to an interval scale. In hemophilia, there is the absence of application of RMT as a filter for patient-centric instrument choice and as a basis for meeting gaps in outcomes requirements [7]. This is an oversight that should be corrected. Unless we have confidence in the interval or ratio properties of our response assessment, we are locked into recommendations in instrument choice that lack merit; or at least a coherent justification for their choice as unidimensional or single attribute interval scales. By default, unless for subjective or patient centric instruments we can apply the filter of Rasch assessment all scores from these instruments are ordinal and cannot capture response to therapy.

IMPOSSIBLE UTILITY SCORES AND QALYs

The emphasis in HTA on preference scores and QALYs is in large part the result of the early commitment to incremental simulated cost-per-QALY claims; the rejection of hypothesis testing in favor of approximate information [11]. To create the assumption driven simulation model you require a preference score; what is overlooked is that the utility or preference score must have bounded ratio scale properties; that is a true zero, invariance of comparison and capped at unity. None of the existing utility or preference techniques (e.g., standard gamble [SG] or time trade off [TTO] or the EQ-5D-3L/5L, HUIMk2/3, SF-6D or SF-36 [PCS and MCS components] have this property. They produce, as they are defining or describing limited descriptions of health states, only composite ordinal scores; they lack dimensional homogeneity and construct validity [3]. As an ordinal score they cannot support standard arithmetic operations (addition, subtraction, multiplication, division); only non-parametric operations which exclude claims for mean values and measures of dispersion. At best we can report medians and an interquartile range. As the QALY requires time to be multiplied by a preference score, the QALY is mathematically impossible [12]. The preference score must have ratio properties.

Judged against the required standards for patient outcomes in terms of RMT, we have a long way to go in hemophilia value claims to ensure that patient reported outcomes are being appropriately measured. These utility instruments lack a fundamental requirement of RMT, conjoint simultaneous measurement. An example would be needs fulfillment: what are the needs (in holistic terms) of adults, adolescents and caregivers, with a target group defined by disease severity in hemophilia, given the difficulty of the need and the ability of the respondent to realize that need, in evaluating how they respond to competing therapy interventions, including the option of gene therapy. While there are no instruments in hemophilia which addresses this latent construct, there are many examples of how instruments for these groups can be developed to meet Rasch standards. A recent example in the Alzheimer Patient Partner Life Impact (APPLIQue) questionnaire, a needs fulfillment instrument for caregiver quality of life which has the required Rasch interval measurement properties and can be translated to a bounded ratio scale [13] [14] [15] . Other RMT questionnaire examples, with particular relevance to rare disease, are plexiform neurofibromas, ankylosing spondylitis and systemic lupus erythematosus [16] [17] [18].

The point to emphasize is the false commitment to assumption driven cost-per-QALY simulations where the exercise assumes these instrument scores have implicit (i.e., assumed) bounded ratio properties. ICER is not alone in presenting assumption driven cost-effectiveness claims for hemophilia therapies, including gene therapy; studies keep on surfacing, producing different non-evaluable claims for the same therapy areas and, presumably, will continue to do so, even as marketing exercises [19] [20] . They all rely on Markov modeling or similar assumption driven lifetime simulations to create imaginary claims. A common feature, as with ICER, is to base the claims on incremental cost-utility models; there is no perception of the limitations of fundamental measurement in the application of ordinal preference score and the impossibility of the QALY as a construct. The fact that the value claim for cost-effectiveness is not empirically meaningful does not appear to be an issue

Systematic reviews of these various modeled hemophilia studies also fail to recognize the limitations imposed by fundamental measurement and the importance of value claims that are credible, evaluable and replicable [21]. Pointing to what is seen as deficiencies in assumption driven simulations in hemophilia, to include overlooking heterogeneity in the patient population, the absence of treatment strategies and

DOI: https://doi.org/10.24926/iip.v13i4.5111

inadequate control for confounders, the authors propose a set of 12 standards, based on the CHEERS format, to make the modeled simulations with their imaginary cost effectiveness claims more helpful to decision makers [22]. These include basing the evaluation on systematic reviews, assessment over the lifetime of a target group, using a generalized measure of benefit (i.e., QALYs) and to characterize uncertainty through probabilistic sensitivity analysis. Whether these CHEERS plus standards can be considered a basis for an imaginary assumption driven simulation being 'more helpful' is a moot point as there is no interest expressed in the standards of normal science and the limitations of fundamental measurement with the application of RMT. A subsequent systematic review merely reinforces this commitment to non-science in value claims although admitting that methodological variations among the more recent studies and difference in treatment schemes made comparative assessments difficult [23]. At least, with this diversity, decision makers will not be spared loss of choice in these 'helpful' imaginary cost-effectiveness claims, although there is the perennial concern with bias and the reverse engineering of imaginary claims in hemophilia and across the board in disease states [24].

The problem with arbitrary assumptions is that there can be no claim that one bundle of assumptions regarding future modeling for cost-effectiveness can be claimed to be superior (or 'more realistic') than another. This is a trivial point, but one that is typically ignored. Recent systematic reviews of cost-utility assessment of hemophilia therapies have found, as expected, considerable variation in modeled claims. Thorat et al, for example, in a further systematic review of models to support claims for the overall value of hemophilia treatment based on cost-utility analyses found that cost-effectiveness claims for treatments *varied widely based on variations in study design, including differences in time horizon, discount rates and medical interventions* [25]; a conclusion that echoes an earlier systematic review by Valente et al concluding that*: These studies reported remarkably different results, using utility values based on different assumptions and data sources* [26] .

These results are entirely expected; they follow from the commitment in traditional HTA to the construction of modeled approximate information to support formulary decisions. Hypothesis testing has been put to one side in favor of assumption-based evidence to support cost-effectiveness claims and formulary decisions; this means there is no coherent basis for empirically assessing competing claims as there would be with hypothesis testing with empirical claims. A denial, therefore, of the standards of normal science and progress driven by the process of conjecture and refutation. We can forget attempts to impose standards for modeling; the claim that we have reference case standards is too open ended to be of any use given, for example, the opportunity to apply a range of utility scores and even reverse engineer the modeled claims [27]. The result is what may be described as an 'open season' for imaginary cost-utility models in hemophilia; none of which should be taken seriously by formulary committees. This means rejecting the notion of imaginary blanket cost-effectiveness claims and adopting a single attribute, protocol driven value assessment framework for specific product claims in hemophilia, whether these are in clinical, PRO quality of life or drug and resource utilization terms.

RASCH, PROMIS AND ITEM RESPONSE THEORY

Item response theory (IRT) and the Patient Reported Outcomes Measurement Response System (PROMIS) item bank and proposed item sets are of limited use in target patient population outcomes assessment. In the case of IRT and PROMIS, various items and instruments are not designed to produce interval scores, with the possibility of a translation to a ratio score. Certainly, there is a potential application of the PROMIS item bank as a key part of the recommended instruments for the health outcomes set in hemophilia, linked to hemophilia specific items (van Balen et al, Table 3 [5]) and the application in hemophilia target groups of T-statistic distributions, including capturing individual claims for response to therapy; but this is as far as we can go for both dichotomous and polytomous PROMIS item or IRT constructs until RMT required standards are applied..

The reason for this limited application, judged from a Rasch perspective, is where the Rasch model should not be considered as the one-parameter logistic model (1PL IRT), as the approach to conceptualizing the relationship between data and theory. IRT, in common with Classical Test Theory (CTT), is focused on fitting a prior model to the available data to create items; the objective being to report the best fit even if this involves the introduction of supplementary model parameters. Rash is conceptually and paradigmatically distinct with the primary focus on meeting standards for fundamental measurement where data items have to meet the Rasch model specification to support reasonable interval-measured conclusions and inferences from the data. Data collection, the choice of items, must be guided by Rasch considerations. This points to the distinction between fitting a model to the data and, in the Rasch approach, selecting data items to fit the model. This distinction is important because it points to a fundamental difference between IRT and Rasch measurement. Thus, while PROMIS/IRT models are 'exploratory and descriptive'; Rasch models are 'confirmatory and predictive' where the data are required to confirm the model. The focus in Rasch modeling is on the size and structure of residuals to confirm that the principles of conjoint simultaneous measurement have been realized sufficiently to justify the claim that the measure of a latent construct has invariant interval properties. This applies to both dichotomous and polytomous instruments where the requirement is to calibrate instruments. We are not concerned as with IRT to fit the model to data to minimize residual variance. This does not mean we drop items in the Rasch model fitting willy-nilly, but attempt to find out, in terms of the latent construct, why an item is apparently redundant. The bottom line, as Andrich would argue, is that the

DOI: https://doi.org/10.24926/iip.v13i4.5111

Rasch and PROMIS/IRT approaches are incompatible paradigms [28] [29]. The Rasch measurement model is the suitable framework if our objective is fundamental measurement.

If the Rasch model is taken as the exemplar framework for both dichotomous and polytomous PROs, the question that can then be addressed is the extent to which PROMIS or IRT based instruments fall short when evaluated in terms of the criteria that RMT requires to judge whether or not the PROMIS item-selected instrument meets Rasch requirements; in particular in respect of item-based polytomous instruments. While there is no hard and fast template for conducting a Rasch analysis of existing polytomous instruments, the following elements should be part (and are supported by assessment software such as RUMM2030, WINSTEPS and R) of an assessment to evaluate whether or not the instrument has properties that may be considered an acceptable approximation to an interval score [30].

- Overall instrument and item functioning
- Unidimensionality of underlying construct
- Local independence of items
- Category and threshold functioning
- Differential item functioning
- Person and item alignment

While the extant literature is limited, two recent studies could be trail setters in this *ex post facto* assessment of PROMIS type instruments in the cases of depression and fatigue [31] [32] . These assessment address, respectively, the IRT 28 and 51 item depression scales and the IRT fatigue scale.; both of these are part of the core PRO set proposed for hemophilia. The results of these two analyses are not encouraging. While there is no magic threshold given the criteria and the degree of overall subjectivity involved, the results for depression point to a number of significant deficiencies involving item fit, failure to capture the latent trait of interest in a sample of responses to scale targeting (e.g., for those with lower depression) which were not resolved with the 51- compared to the 28-item scale and the absence of a single construct. While the depression scales might be improved by item re-assessment, the scales were unsatisfactory, even to considering additional conceptual categories of depression. The fatigue scale assessments were broader involving the FACIT-F scale, the SF-36 vitality scale and the LupusQoL Fatigue instrument. The results challenged the use of all three instruments; they were not considered suitable for clinical assessments. Item fit was poor, even when a selection of the best-performing items was made. The FACIT-F total score was considered inadequate with multiple underpinning concepts. The SF-36 vitality and LupusQol Fatigue scales failed to address relevant fatigue issues leading to a lack of precision in the scores.

On the positive side there is now a growing literature, notably in the area of rehabilitation medicine, that has addressed the issue of establishing Rasch standards for instruments and the evaluation of existing instruments in target patient populations or disease areas. Most recently the publication of the Rasch Reporting Guideline for Rehabilitation Research (RULER) that is intended to ensure authors, reviewers and editors have uniform guidance about how to write and evaluate research on rehabilitation outcomes assessments in terms of six psychometric Rasch domains [33] [34]. These include conceptual content (e.g., unidimensionality; structural validity); rating scale step structure; overall model and individual item fit; and measurement invariance, reproducibility and reliability for differential item functioning.

RASCH AND HEMOPHILIA OUTCOMES

Once we accept the need to meet the standards of normal science, including fundamental measurement, we have to accept the imperative of a RMT 'filter' to accept value claims. Clearly, there will be clinical value claims (e.g., bleed rates) which will meet ratio properties, as well as claims for drug and resource utilization. It is the subjective latent construct PRO claims that present a challenge. These must meet Rasch measurement standards. This requirement is, unfortunately absent from the criteria applied in outcome measures by ICHOM and others, despite their apparent commitment to a program of long-term research in hemophilia. Meeting the standards of fundamental measurement is a necessary precondition to therapy evaluations, it must rest, in PRO claims, on a sound theoretical foundation; one that is missing in non-evaluable ordinal cost-utility claims instruments. There is an important distinction between the pragmatic *ex post facto* application of the Rasch model and a commitment to Rasch in the process of instrument development. This supports the role that Rasch measurement can and must play in instrument development.

Patient centric value claims in hemophilia face a critical hurdle: do they meet the required measurement properties, either ratio or interval. The key question: what are we trying to measure in hemophilia? If we are to develop items and their selection then we must understand the nature of the possible latent traits and whether they can be measured following RMT. Those proposing core measures in hemophilia have a long way to go. There are a number of examples of measures which have unknown measurement properties and for which, as a default position, we must accept as only ordinal scores and which have no role in evaluating response to therapy.

Likert-based polytomous instruments in hemophilia for which we must assume only ordinal properties include: (i) the Haem-A-QoL (adult) and Haemo-QoL (pediatric) instruments which rely on the summation of integer scores from Likert response scales (some reordered) with subscales and an overall score to create composite total raw and transformed ordinal scores [35]; (ii) the Canadian Hemophilia Outcomes-Kids Life Assessment Tool (CHO-KLAT) version 2.0 (youth), with a recent revision to version 3.0 (youth) with 40 items and seven domains with Likert integer summed sub-scales and total score standardized to 100 [36]; and (iii) the Hemophilia Activities List (HAL) designed to

DOI: https://doi.org/10.24926/iip.v13i4.5111

measure the impact of hemophilia on self-perceived functional abilities in adults with hemophilia has 42 Likert items in seven domains where the most important outcomes are overall integer sum score and three component scores for upper extremity activities, basic lower extremity activities and complex lower extremity activities, with each domain also scored and, more recently, a shortened version of 18 items [37]. None of these instruments has the required interval scale measurement property for a single attribute which is the unique contribution from applying Rasch rules to instrument development. These various instruments only produce ordinal scores or counts and are illegitimately used to claim therapy response despite considerable attention given to what the various authors have judged the appropriate psychometric tests for these instruments over some 20 years; the question of fundamental measurement was never addressed.

PROTOCOLS

Value claims for hemophilia therapies should be accompanied by a protocol detailing how the claim is to be assessed; this applies equally to patient-centric outcomes as it would to purely clinical, drug and resource utilization claims. There is a need for meeting the standards for fundamental measurement and the application of Rasch modeling in hemophilia outcomes, in order to establish a real-world evidence basis for therapy choice and the basis for ongoing research programs and even outcomes-based contracting. Judged from this perspective, assumption driven simulations are insufficient for sound decision-making. There is no basis for progress in our understanding of the benefits of competing therapies or, what Popper would describe as the evolution of our objective knowledge in hemophilia [38]. This applies not only to patient reported outcomes to include quality of life of patients and caregivers, but to purely clinical claims defined in ratio terms and those for drug and resource utilization.

CONCLUSIONS

If we are to accept the relevance of patient centric instruments in hemophilia then we have to recognize the importance of RMT. As stated by Wright, and quoted by Bond, *Rasch models are the only laws of quantification that define objective measurement, determine what is measurable, decide which data are useful, and expose which data are not* [6]. There are minimum standards for measurement which must apply for both dichotomous and polytomous hemophilia instruments irrespective of the target patient population. If the existing instruments cannot be defended then new instruments must be devised to meet Rasch standards (e.g., needs fulfillment quality of life). Until we achieve these standards, the absence of acceptable measurement will prejudice therapy response claims and proposals for ongoing research strategies for new therapies, including pathbreaking gene therapies in hemophilia.

Conflicts of Interest: PCL is an Advisory Board Member and Consultant to the Institute for Patient Access and Affordability, a program of Patients Rising

Note: The opinions contained in this paper are those of the author (PCL).

DOI: https://doi.org/10.24926/iip.v13i4.5111

REFERENCES

[1] Tice J, Walton S, Herce-Hagiwara B et al. Gene Therapy for Hemophilia B and An Update on Gene Therapy for Hemophilia A: Effectiveness and Value; Evidence Report. Institute for Clinical and Economic Review, November 2, 2022.

[2] Langley P. Nothing to Cheer About: Endorsing Imaginary Economic Evaluations and Value Claims with CHEERS 22 [version 1; peer review: peer reviewed; 2 approved] *F1000Research* 2022, 11:248

[3] Langley PC and McKenna SP. Measurement, modeling and QALYs [version 1; peer review: 2 approved] *F1000Research* 2020, 9:1048 https://doi.org/10.12688/f1000research.25039.1

[4] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-2023 value assessment framework for constructing imaginary worlds. *Inov Pharm*. 2020;11(1):No. 12

[5] Andrich D, Marais I. A Course in Rasch Measurement: Measuring in the Educational, Social and Health Sciences. Singapore: Springer, 2019

[6] Bond T, Yan Z, Heene M. Applying the Rasch Model: Fundamental measurement in the human sciences (4th Ed.) New York: Routledge, 2021

[7] Van Balen E, O'Mahoney B, Cnossen M et al. Patient-relevant health outcomes for hemophilia care: Development of an international standard outcomes set. *Res Pract Thrombosis Haemostasis*. 2021;5:e12488

[8] Hunt S, McKenna S, McEwen J et al. The Nottingham Health Profile" subjective health status and medical consultations. *Soc Sci Med A*. 1981;15(3 Pt 1):221-9

[9] Wright B, Linacre J. Observations are always ordinal; Measurements, however, must be interval. *Arch Phys Med Rehab.* 1989; 70(12):857-60

[10] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960

[11] Neumann P, Willke R, Garrison L: A Health Economics Approach to US Value Assessment Frameworks – Introduction: An ISPOR Special Task Force Report. *ValueHealth.* 2018; **21**: 119–123.

[12] Langley P. The Great I-QALY Disaster. *InovPharm*. 2020; 11(3): No 7

[13] Hagell P, Rouse M, McKenna S. Measuring the impact of caring for a spouse with Alzheimer's disease: Validation of the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue). *J App Measurement*. 2018;19(3):271-282

[14] McKenna S, Rouse M, Heaney A et al. International development of the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue). *Am J Alzheimer's Dis and Other Dementias*. 2020;35:1-11

[15] Langley P, McKenna S. Fundamental measurement: The Need Fulfillment Quality of Life (N-QOL) Measure. *InovPharm*. 2021;11(3): no. 17

[16] Heaney A, Wilburn J, Rouse M et al. The development of the PlexiQoL: A patient-reported outcome measure foe adults with neurofibromatosis type 1 - associated plexiform neurofibromas. *Genet Genomic Med*. 2020;8(12):e1530

[17] Heaney A, McKenna S, Hagell P et al. Improving scoring precision and internal construct validity of the Bath Ankylosing Spondylitis Disease Activity Index using Rasch Measurement Theory. *J Rheumatol*. 2020;47(3):354-61

[18] Doward LC, McKenna SP, Whalley D, Tennant A, Griffiths B, Emery P, et al. The development of the L-QoL: a quality-of-life instrument specific to systemic lupus erythematosus. *Ann Rheum Dis.* 2009;68(2):196–200.

[19] Bolous N, Chen Y, Wang H et al. The cost-effectiveness of gene therapy for severe hemophilia B: A microsimulation study from the United States perspective. *Blood*. 2021; 138(18):1677-90

[20] Cook K, Forbes S, Adamski K et al. Assessing the potential cost-effectiveness of a gene therapy for hemophilia A. *J Med Econ*. 2020; 23(5):501-12

[21] Drummond M, Houwing N, Slothus U et al. Making economic evaluations more helpful for treatment choices in hemophilia. *Haemophilia*. 2017;23(2):e58-e66

[22] Husereau D, Drummond M, Petrou S et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) – explanation and elaboration: A report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *ValueHealth*. 2013;16:231-50

[23] Cortesi P, D'Angiolella L, Lafranconi A et al. Modern treatments of haemophilia: Review of cost-effectiveness analyses and future directions. *Pharmacoeconomics*. 2018;36(3):263-284

[24] Langley P. Facilitating bias in cost-effectiveness analysis: CHEERS 2022 and the creation of assumption-driven imaginary value claims in health technology assessment [version 1; peer review: 2 approved]. *F1000Research* 2022, 11:993

[25] Thorat T, Neumann P, Chambers J. Hemophilia burden of disease; A systematic review2 of the cost utility literature for hemophilia. *J Man Care Specialty Pharmacy*. 2018;24(7):632-42

[26] Valente M, Cortesi P, Lassandro G et al. Health economic models in hemophilia A and utility assumptions from a clinician's perspective. *Pediatr Blood Cancer*. 2015; 62(10):1826-31

[27] Langley P. Let a Thousand Models Bloom: ICER Analytics Opens the Floodgates to Cloud Pseudoscience. *Inov Pharm*. 2021;12(1):No. 5

[28] Andrich D. Controversy and the Rasch Model: A characteristic of incompatible paradigms? *Med Care*. 2004;42(1 Suppl):17

[29] Andrich D. Rating scales and Rasch Measurement. *Expert Rev Pharmacoecon Outcomes Res*. 2011;11(5):571-85

[30] Combrinck C. Is this a useful instrument? An introduction to Rasch measurement models, in Kramer S et al (eds.) Online Readings in Research Methods. Psychological Society of South Africa. Johannesburg, 2020

[31] Cleanthous S, Barbic S, Smith S et al. Psychometric performance of the PROMIS depression item bank: a comparison of the 28 and 51 item versions using Rasch measurement theory. *J Patient-Reported Outcomes*. 2019;3:47

[32] Cleanthous S, Bongardt S, Marquis P et al. Psychometric analysis from EMBODY1 and 2 clinical trials to help select suitable PRO fatigue scales for future systemic lupus erythematosus studies. *Rheumatol Ther*. 2021;8:1287-1301

[33] Mallinson T, Kozlowski A, Johnston M et al. Rasch Reporting Guidelines for Rehabilitation Research (RULER): The RULER Statement. *Arch Physical Medicine Rehab*. 2022;103:1477-86

[34] Van de Winckel A, Kozlowski A, Johnston M et al. Reporting guideline for RULER: Rasch Reporting Guideline for Rehabilitation Research: Explanation and Elaboration. *Arch Physical Medicine Rehab*. 2022;103:1487-98

[35] Wyrwich K, Krishnan S, Poon J et al. Interpreting important health-related quality of life change using the Haem-A-Qol. *Haemophilia*. 2015; 21:578-84

[36] Price V, Dover S, Blanchette V et al. Updating the Canadian Hemophilia Outcomes-Kids Life Assessment Tool (CHO-KLAT) in the era of extended half-life clotting factor concentrates. *Res Pract Thromb Haemost*. 2021;5:403-11

[37] Kuijlaara I, van der Net J, Buckner T et al. Shortening the Haemophilia Activities List (HAL) from 42 items to 18 items. *Haemophilia*. 2021;27:1062-1070

[38] Magee B. Popper. London: Fontana, 1974