Addressing traceability in social measurement: establishing a common metric for dependence

T Salzberger¹, S Cano², L Abetz-Webb³, E Afolalu⁴, C Chrea⁴, R Weitkunat⁴, K Fagerström⁵, J Rose⁶

¹ WU Wien, University of Economics and Business, Vienna, Austria, thomas.salzberger@wu.ac.at
² Modus Outcomes, Spirella Building, Letchworth Garden City, UK, stefan.cano@modusoutcomes.com
³ Patient-Centered Outcomes Assessments, LI, Macclesfield Cheshire, UK, linda.abetz-webb@p-coa.com
⁴ PMI R&D, Philip Morris Products S.A., Neuchâtel, Switzerland, Esther.Afolalu@pmi.com, Christelle.Chrea@pmi.com, Rolf.Weitkunat@pmi.com
⁵ K. Fagerström Consulting AB, Vaxholm, Sweden, karl.fagerstrom@hotmail.com
⁶ Rose Research Center, Raleigh, NC, USA, jed.rose@roseresearchcenter.com

E-mail: thomas.salzberger@wu.ac.at

Abstract. Measurement in the social sciences is characterised by a multitude of incompatible paradigms, most of which fail to adhere to fundamental principles of measurement explicated in metrology. This has led to a fragmentation of instruments that are hard to interpret and lack a common reference. Rasch measurement theory, the metrological framework of the Rasch model, combined with strong substantive theories of the measurand, has the potential to advance measurement in the social sciences significantly. The Rasch model establishes a reference standard with a common unit of measurement against which different instruments can be calibrated ensuring traceability in social measurement. Uncertainty in the calibration process is addressed by a standard error of measurement for a specific estimate taking into account that precision varies over the continuum of the measurand. A strong substantive theory of the measurand allows for concrete predictions that can be tested empirically with positive evidence supporting accuracy. We exemplify the potential of the Rasch model by developing the ABOUT-D dependence, a new instrument for the measurement of dependence on tobacco-containing products (TNP). The instrument provides comparable measurements of dependence on different TNP's as well as dependence on multiple TNP's used concurrently. A co-calibration of the new instrument and existing, product-specific instruments allows for a conversion of scores of existing instruments into scores on the new instrument, and vice versa, via the established equal-unit reference scale. The score comparability adds to traceability in the measurement of dependence, is likely to facilitate the acceptance of the new instrument, and helps avoid a discontinuity in the research of dependence.

1. Introduction
Measurement is a precisely defined concept, with traceability and a stated range of uncertainty being its core properties [1, 2]. While these principles apply to all disciplines alike [3], measurement in the social
sciences (also referred to as social, or human, measurement) has been plagued by a multitude of incompatible paradigms [4, 5] bearing on very different definitions of measurement [6]. These paradigms have contributed to the widespread view that social measurement and its properties are the subject of negotiation and pragmatism. Numbers resulting from coded observations (such as ticking a box in a questionnaire) have been interpreted as measurement with little, if any, justification. Most paradigms fail to reveal anomalies disproving successful measurement implying that the scientific mechanism of self-correction by replication stops functioning. Measurement models are supposed to best describe the data rather than expose anomalies revealing weaknesses in the measurement process. Furthermore, concepts to be measured are often vaguely defined, leaving plenty of room to substantiate them after the fact, essentially defining them in terms of the items retained in the instrument, quite often according to arbitrary statistical criteria. As a consequence, instruments tend to be fragmented and weakly linked to one another. To attain the same level of scientific rigour as the natural sciences with respect to measurement, the social sciences need to reconsider the underpinnings of measurement. The paradigm of Rasch measurement [7], or Rasch measurement theory (RMT [8, 9]), lends itself as a conceptualisation, or philosophy, of measurement that takes traceability and uncertainty adequately into account. The measurement of dependence on tobacco- and nicotine-containing products (TNP) illustrates how RMT contributes to concept-driven measurement in the social sciences, providing interpretable measures with a specified level of uncertainty. In the social sciences, the latter is typically addressed by the standard error of measurement based on reliability.

2. A measurement theory for the social sciences

Measurement instruments in the social sciences typically consist of a set of items (questions or statements) collectively supposed to represent the measurand. Each item is a reference to which subjects can be compared. Each response of a subject to an item is the manifest outcome of such a comparison as the measurement method. The probability of a particular response is modelled as a function of the item’s quantity and the subject’s quantity, which are represented by an item location and a person location parameter, respectively. Item and subject quantities are estimated from a series of observed responses to a set of items forming a scale representing the same measurand. In order to provide measurement, items need to be like stable markings on a ruler that are invariant with respect to the respondents. The independence of item locations from the specific respondents, instrumental in the item location estimation, is achieved by the Rasch model for measurement [7] by virtue of its defining characteristic of specific objectivity [10, 11]. In the case of binary response formats (dichotomous response scales, such as yes versus no or agree versus disagree, scored 0 and 1, respectively), one item location parameter is estimated. Ordered multi-categorical formats (polytomous response scales, such as rating scales) require the estimation of a set of threshold locations marking the boundaries between adjacent response categories. The expected score on an item as a function of the item location and the person location forms an s-shaped curve, represented by a logistic function. In the Rasch model, there is no additional parameter describing variation in the slope of this curve. Thus, in the dichotomous case, the Rasch model requires the slope to be the same for all items to retain an equal unit of measurement across the continuum. In the polytomous case, the slope at each threshold has to be the same across all thresholds and items. This implies that respondents discriminate equally between each pair of adjacent response categories.

Valid measurement requires two conditions to hold. First, the observed responses need to correspond to the responses expected under the model. Second, the item location estimates need to form a meaningful hierarchy in terms of the measurand and its conceptualisation. The first requirement is addressed by various statistical tests of fit assessing whether the observed responses match the responses expected under the model. Another approach to test whether the data fit the model is based on specific objectivity. It implies that item parameters are invariant regardless of the subjects used to estimate them. Thus, item location estimates must not differ between different respondent groups. All of these tests involve formal statistical procedures that, by their very nature, do not take item content into account. Furthermore, they are based on empirically estimated item location parameters with no reference to
predictions by the conceptual theory of the measurand. Thus, mere model fit is a necessary, but not a sufficient condition of measurement.

Therefore, to substantiate accuracy of measurement, the estimated item parameters have to be linked to the conceptual theory. To begin with, the item hierarchy in terms of item locations needs to be meaningful and interpretable. It demonstrates what more or less of the property assessed actually means, allowing for an unambiguous interpretation of person measures. If the conceptual theory proposes a particular item hierarchy, comparing the empirical item hierarchy and the expected hierarchy tests whether theoretical predictions apply. Ideally, the theory predicts a concrete location for each item based on some exposed mechanism driving the item’s location [12]. While successful implementations of such elaborate construct theories are still rare, the focus in RMT has shifted towards more powerful construct theories.

In the case of less well-developed conceptual theories, empirical insights in terms of the actual item hierarchy revealed by the Rasch model may help advance the conceptual theory. The contribution of applications of the Rasch model to theory extension is particularly relevant in the transition phase, which has to cope with two obstacles. First, an abrupt change from firmly established but inadequate rules to new principles is likely to face resistance [13], and existing instruments will continue to be defended and will not fall into sudden disuse. Second, an abrupt shift may lead to a discontinuity in substantive research, implying unjustified devaluation of existing research findings. Rather, legacy instruments should be scrutinised, re-evaluated, improved when needed, and linked to new measures and to one another. In this regard, RMT is particularly powerful, allowing for establishment of a common metric for different measures of the same measurand, which is an important contribution to traceability.

3. The measurement of dependence on TNPs

Nicotine dependence is a key driver underlying maintenance of use of TNPs. The measurement of nicotine dependence has primarily focused on assessing product-specific dependence (notably cigarettes) rather than overall dependence on TNPs [14]. A broad range of alternative TNPs have become available (e.g. electronic cigarettes, heat-not-burn tobacco products, smokeless tobacco), and some of these products are currently under investigation by regulatory bodies as part of a population harm reduction strategy [15-17]. The measurement of dependence is currently hampered by two obstacles: Measures for dependence on different TNPs are not comparable; and no measure for dependence on multiple TNPs is available [18]. To overcome these limitations, a new instrument (the ABOUT-Dependence, as part of the ABOUT Toolbox initiative [19]) was developed.

First, a conceptual model of dependence was generated based on a literature review, expert opinions, and concept elicitation interviews with TNP users (n=40 [20, 21]). Second, a draft version of the instrument comprising 19 potential items was developed, with items best representing the aspects of the concept of interest, response categories meaningful to intended respondents (TNP users), and appropriate recall periods. The items were grouped into seven aspects of dependence (such as compulsion to use, automaticity of using, or self-awareness of dependence). In the cognitive interviews, some of the items were identified as being conceptually redundant (alternative phrasing of the same issue), while one item did seem to be relevant, but only to some of the participants. However, all 19 items were tested psychometrically to identify the best-working items when multiple items capturing the same issue were generated. Third, in a field test, data from a sample of 2,434 respondents in the U.S. comprising 1,253 poly-users and 1,181 exclusive users (seven different TNPs considered: conventional cigarettes, e-cigarettes, smokeless tobacco, cigars, pipes, waterpipes, and nicotine replacement therapy) were collected to assess the proposed instrument. As strict unidimensionality was not tenable, the conceptualisation was reconsidered, and eventually, three domains of dependence were proposed (extent-of-use, covering the urgency or pervasiveness with which the product(s) is/are used; doing, comprising aspects of how dependence impacts on daily activities; and feeling, related to symptoms and signs of dependence experienced by TNP users). With regard to the doing domain, a partial ordering of the items was to be expected, with using more of the product than intended being a basic symptom of dependence that should be endorsed most easily. Endorsing stopping an activity in order to use a TNP
was supposed to require less dependence compared to *avoiding an activity altogether*. For the *feeling* domain, a *strong desire* was supposed to be more readily endorsed than finding it hard to control the *urge*, as the latter implies the former but not vice versa.

The application of the polytomous Rasch model confirmed the three domains of the revised conceptualisation, revealed redundancy where it had been expected, and identified one item suspected to misfit as inadequate. The reduction of the number of items to 12 by omitting six redundant items and one misfitting item did not result in any lack of content coverage. The partially expected ordering of items was confirmed, and the full item hierarchy turned out to be conceptually meaningful. As the three domain measures were correlated in the order of 0.5 to 0.8 in the entire sample, a composite score capturing overall dependence was proposed covering all three domains. The *extent-of-use* domain required an adjustment for differential item functioning by the type of TNP used.

Technically, the composite scores were analysed by applying the Rasch model to domain sub-scores as super-items (or subtests). The Rasch model transformed the total sum score (raw score) into a linear measure on a logit scale, which was then linearly transformed into a more convenient and accessible metric of 0 to 100. By ordering all threshold locations across all three domains, for each observed total score, the most likely score on each domain was inferred. Subsequently, each domain score was characterised by the most likely response pattern based on the previous application of the Rasch model to the domain items. By doing so, a comprehensive interpretation of each total score on the composite measure was achieved.

After successfully establishing a meaningful and interpretable structure of items forming composite scores, various existing instruments can be linked to the newly established metric. This is done by co-calibrating the new instrument (three domains as subtests) and existing instruments (represented by another subtest), whereby the subtest parameter estimates for the new instrument were anchored to values estimated in preceding analysis of the new instrument only. This procedure provides the opportunity to transform raw scores on various existing product-specific instruments to a common metric established by the new instrument. Figure 1 shows how ABOUT-Dependence measurements among exclusive cigarette users are converted to measurements by the very frequently used Fagerström Test for Nicotine Dependence (FTND) [22], recently renamed the Fagerström Test for Cigarette Dependence [23], and vice versa. The item locations for each instrument are represented by squared dots on the common linear metric. Each raw score on the FTND matches a score on the ABOUT-Dependence. For example, a score of 2 on the FTND is equivalent to a score of 12 on the ABOUT-Dependence. Conversely, scores of 10 to 14 on the ABOUT-Dependence all map onto a score of 2 on the FTND. This shows that the new instrument has a better resolution compared to the FTND.

![Figure 1](attachment:figure1.png)

**Figure 1.** Conversion of the ABOUT-Dependence raw score (composite score) into the FTND raw score via a common linear metric.

Figure 2 illustrates the conversion of ABOUT-Dependence raw scores to linear measurements. Each raw score from 0 to 50, on the x-axis, is transformed into a linear measurement of dependence, on the
y-axis. The transformation is non-linear demonstrating the raw score is not a linear measure. The error bars represent 95% confidence intervals for the estimated measures illustrating measurement uncertainty. Towards the extremes, measurement uncertainty increases, while most precise measures are provided in the middle of the scale.

**Figure 2.** Conversion of the ABOUT-Dependence raw score (composite score) into a linear measurement with 95% confidence intervals for measurement uncertainty.

### 4. Conclusions

The measurement of TNP dependence has focused on specific tobacco products, most notably cigarettes. Today, such instruments are no longer fit-for-purpose due to the diversity of products available and the prevalent habits of using multiple products concurrently. This is an obstacle for health policy makers, regulatory bodies, and researchers in the field of tobacco regulatory science as dependence across different TNPs is hard to compare. Based on a conceptual model of dependence, the interplay of insights from the psychometric analysis applying the Rasch model and further conceptual considerations, a proposed three-domain structure of dependence could be empirically supported. The high correlations of these three domain measures justified the formation of a composite score summarising dependence by one overall measure. Compared to existing measures, the ABOUT-Dependence provides a better interpretable measure of dependence by referring to individual item and response category locations. In addition, the ABOUT-Dependence provides measures with better resolution (Figure 1) compared to existing instruments, such as the FTND. The composite score allows for a consolidation of existing instruments, and the ABOUT-Dependence by providing crosswalks from the metric of one instrument to the metric of another. The transformation of scores from existing instruments to the newly established metric should be interpreted with caution, however, as existing measures do not necessarily meet the requirements of the Rasch model. What is more, the transformed scores retain the level of the original uncertainty, which is typically considerably higher than uncertainty associated with ABOUT-Dependence measurements. When comparing dependence measures across products, it is important to realise that what the instrument measures is the existing dependence on a specific TNP or, in case of poly-users, a bundle of TNPs. In a cross-sectional study, the measurements do not allow for inferring a TNP’s potential to induce nicotine dependence. Applied research, in particular longitudinal studies, will reveal how dependence develops over time and the predictive validity of the instrument.

**References**

[1] JCGM 2012 *JCGM 200:2012 International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM)* 3rd ed (2008 version with minor corrections) (Sèvres: Joint Committee for Guides in Metrology) (www.bipm.org/en/publications-guides/vim.html)
[2] JCGM 2008 JCGM 100:2008 Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement (GUM 1995 with minor corrections) (Joint Committee for Guides in Metrology)

[3] Pendrill L R and Fisher Jr W P 2013 Quantifying human response: linking metrological and psychometric characterisations of man as a measurement instrument J. Phys.: Conf. Ser. 459 1 01205

[4] Cano S et al 2016 On Trial: the Compatibility of Measurement in the Physical and Social Sciences J. Phys.: Conf. Ser. 772 012025

[5] Andrich D 2004 Controversy and the Rasch model: a characteristic of incompatible paradigms? Med. Care 17-I16

[6] Michell J 1997 Quantitative Science and the Definition of Measurement in Psychology Brit. J. Psychol. 88 3 355-383

[7] Rasch G 1960/1980 Probabilistic Models for Some Intelligence and Attainment Tests (Copenhagen: Danish Institute for Educational Research; expanded edition 1980 Chicago, IL: The University of Chicago Press)

[8] Andrich D 2018 Advances in social measurement: A Rasch measurement theory. In F Guillemin, A Leplège, S Briançon, E Spitz, and J Coste (Eds) Perceived Health and Adaptation in Chronic Disease (London: Routledge) pp 66–91

[9] Cano S, Pendrill L R, Melin J, Köbe T and Fisher Jr W P 2018 Metrology for the Social Sciences: A case for Rasch Measurement Theory (not Rasch Analysis) (New York: IOMW)

[10] Rasch G 1977 On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements Danish Yearbook of Philosophy 14 58-93

[11] Fischer G H 1995 Derivations of the Rasch Model Rasch Models, Foundations Recent Developments, and Applications, ed G H Fischer and I W Molenaar (New York: Springer) 15-38

[12] Stenner A J, Fisher Jr W P, Stone M H and Burdick D S 2013 Causal Rasch models Front. Psychol. 4 1-14

[13] Andrich D 2002 Implications and applications of modern test theory in the context of outcomes based education Stud. Educ. Eval. 28 2 103-121

[14] Strong D R et al. 2017 Indicators of dependence for different types of tobacco product users: Descriptive findings from Wave 1 (2013–2014) of the Population Assessment of Tobacco and Health (PATH) study Drug Alcohol Depen. 178 257-266

[15] Gostin L O 2009 FDA regulation of tobacco: politics, law, and the public’s health Jama 302 13 1459-60

[16] Warner K E and Schroeder S A 2017 FDA’s innovative plan to address the enormous toll of smoking Jama 318 18 1755-56

[17] Fagerström K O and Bridgman K 2014 Tobacco harm reduction: the need for new products that can compete with cigarettes Addict. Behav. 39 3 507-511

[18] Kasza K A et al. 2017 Tobacco-product use by adults and youths in the United States in 2013 and 2014 New Engl. J. Med. 376(4) 342-353

[19] Chrea C et al. 2018 Developing fit-for-purpose self-report instruments for assessing consumer responses to tobacco and nicotine products: the ABOUT™ Toolbox initiative [version 1; peer review: 2 approved] F1000Research 7:1878 (https://doi.org/10.12688/f1000research.16810.1)

[20] Chrea C et al. 2018 Development of a tobacco and nicotine products dependence instrument. 24th Annual Meeting of the Society for Research on Nicotine and Tobacco (SRNT) (Baltimore, USA)

[21] Chrea C et al. 2018 Development of a fit-for-purpose tobacco and nicotine products dependence instrument ISPOR (The International Society for Pharmacoeconomics and Outcomes Research) (Europe Barcelona, Spain)
[22] Heatherton F, Kozlowski L T, Frecker R C and Fagerström K O 1991 The Fagerström Test for Nicotine Dependence: A revision of the Fagerström Tolerance Questionnaire *Brit. J. Addict.* **86** 1119–27

[23] Fagerström K O 2012 Determinants of tobacco use and renaming the FTND to the Fagerström Test for Cigarette Dependence *Nicotine Tob. Res.* **14** 75–78