Probabilistic Measurement of Non-Physical Constructs During Early Childhood: Epistemological Implications for Advancing Psychosocial Science

N. Bezruczko ¹,², S. S. Fatani ¹,³

¹ Office of Early Childhood Education, 125 S. Clark, Chicago, USA
² Measurement and Evaluation Consulting, 1524 E. 59th Street, Chicago, USA
³ DePaul University, Department of Psychology, Chicago, USA.

Nbezruhzko@msn.com

Abstract. Social researchers commonly compute ordinal raw scores and ratings to quantify human aptitudes, attitudes, and abilities but without a clear understanding of their limitations for scientific knowledge. In this research, common ordinal measures were compared to higher order linear (equal interval) scale measures to clarify implications for objectivity, precision, ontological coherence, and meaningfulness. Raw score gains, residualized raw gains, and linear gains calculated with a Rasch model were compared between Time 1 and Time 2 for observations from two early childhood learning assessments. Comparisons show major inconsistencies between ratings and linear gains. When gain distribution was dense, relatively compact, and initial status near item mid-range, linear measures and ratings were indistinguishable. When Time 1 status was distributed more broadly and magnitude of change variable, ratings were unrelated to linear gain, which emphasizes problematic implications of ordinal measures. Surprisingly, residualized gain scores did not significantly improve ordinal measurement of change. In general, raw scores and ratings may be meaningful in specific samples to establish order and high/low rank, but raw score differences suffer from non-uniform units. Even meaningfulness of sample comparisons, as well as derived proportions and percentages, are seriously affected by rank order distortions and should be avoided.

1. Introduction

A widely recognized limitation on social science advancement is weakly-defined or "soft" measurement commonly associated with ordinal raw scores. Many areas of nonphysical measurement such as psychology, education, sociology, and economics are dominated by ordinal raw scores, which are notable for their weak conformity to fundamental measurement principles such as formal axioms, invariance theorem, and an explicit concatenation operation [1]. Although quantitative differences between weak and strong measurement methods are well understood logically, their different epistemological implications for scientific knowledge have never been demonstrated. In fact, contemporary social researchers express enormous confusion about fundamental measurement probably stemming from a poor understanding of these implications. Fan [2], for example, compared ordinal raw scores and linear measures and incorrectly concluded they are empirically identical, a naïve position then widely supported among social researchers. This confusion has plagued social research for most of 20th century and underlies psychology's increasing
fragmentation into "narrowly defined sub-disciplines, each pursuing objectives in relative isolation from others, p. 197" [3].

Social sciences have generally adapted to their "soft" measurement status with an elaborate and complex quantitative methodology. Development of probabilistic conjoint measurement, however, now renders these raw score methods obsolete [4]. Unfortunately, raw score methods are firmly entrenched in social research and conventional social science fiercely resists scientific innovation. Barrett [5] stated "psychometric dogma . . . is strangling innovation and creativity in areas concerning the investigation of psychological phenomena (p. 82)." Andrich [6] described issues central to this conflict between raw scores and probabilistic paradigms in Kuhnian terms, and he emphasized benefits to social sciences such as specific objectivity, sufficient estimation, and parameter invariance from shifting to objective probabilistic measurement. Nonetheless, current social research practice finds linear probabilistic models (Rasch models) largely restricted to health sciences and educational testing. This paradigm conflict is preventing an overall synthesis of instrumental and information approaches to measurement, which historically has separated physical and nonphysical measurement [7].

This presentation examines epistemological importance of fundamental measurement for psychosocial science by comparing empirical results from raw scores and linear, probabilistic measures when both are applied to describe school children observations. Results from measuring human differences and change are interpreted for their influence on establishing logically coherent, hypothetical conceptual entities central to scientific theories. An understanding of their different implications for knowledge should clarify the potential benefits to social science for shifting from raw scores to objective, linear measurement, as well as promote consolidation of social and physical measurement. Consequently, a new definition of social measurement is proposed that emphasizes role of linear measurement in isolating knowledge about one or more functions inherent to complex systems. Instrumental to this definition is conformity of these functions to regularities mathematically modelled during the measurement process. This knowledge is reproducible, exact, objective, and heuristically dependent on linear measurement operations.

2. Problem

Social researchers commonly assume numerals derived from counting observations represent locations on a real number (equal-interval) scale. In fact, any random selection of journals sponsored by American Psychology Association typically shows empirical results reported not in objective linear units as would be expected of a scientific organization but sample-dependent ordinal scores. Even more troubling are highly touted applications of sophisticated statistical models with only ordinal score properties yet assume fundamental properties. In general, epistemological limitations of raw scores are not widely appreciated among typical social researchers, and this problem is not limited to American researchers. In fact, European social researchers currently lead the claim that axiomatic measurement foundations and particularly probabilistic conjoint additivity are irrelevant to psychology [8]. Consider the following statement, "The family of procedures that scientists . . . regard as instances of measurement is diverse and incoherent and has few universal characteristics, p.1062" [9]. In view of widely acknowledged and long standing foundations for scientific measurement and contemporary emphasis on a universal framework [10], "Can this position be justified?"

Scientific measurement authorities generally agree ordinal quantitative structure cannot describe "How much" observations differ because space between rank positions is unknown. In fact, linear magnitude displacement is a necessary and sufficient condition for measuring change and fundamental to describing human development. Studies have found raw score units differ across a score distribution by multiples of four and five units [11]. Consequently, common operations such as forming ratios between ordinal scores to calculate proportions are not meaningful and common statistical procedures are invalid. Molenaar [12, 13] described problems of population-based statistics for measuring within-subject developmental change.

In addition to numerical properties, objectivity of ordinal raw scores is problematic. Raw scores and derived rank orders are dependent on arbitrary samples, which confound rank position with sample
variability, an invariance violation of fundamental measurement. In principle, raw score objectivity or invariance cannot be maintained across stochastic observation samples.

Finally, raw score-based theory development is problematic because conceptual entities cannot be linked to mathematical abstractions nor integrated and consolidated with more comprehensive scientific theories. Non-uniform units across rank order scales virtually eliminate establishing functional relations with theoretically important variables, which inhibits consolidation of scientific knowledge. Further compounding raw score problems are social researchers who flagrantly violate several measurement axioms and theorems assumed by inferential statistical analyses.

3. Objective
The objective of this presentation is to compare several epistemological criteria of scientific knowledge such as meaningfulness, objectivity, and heuristic value between raw scores or ratings and linear, objective measures when a) describing human differences and b) measuring change. The central question is, "Does empirical knowledge based on ordinal raw scores and ratings versus objective linear measures fundamentally differ?" "If so, do their differences have implications for advancing psychosocial science?"

4. Method
Social researchers have proposed complicated and controversial methods such as raw score gains [14, 15, 16], residualized gains [17], and correlation-based methods [18], which they insist accommodate fundamental limitations of ordinal raw scores. Despite wide use, these raw score adaptations have never been compared with objective, linear methods. In this presentation, raw score sums and Rasch model linear measures of early childhood learning are compared for measurement implications.

4.1. Rasch model foundations
Conjoint probabilistic Rasch models were developed by the Danish mathematician, Georg Rasch, for transforming ordinal raw scores into objective, linear measures [19]. Briefly, Rasch models implement a probabilistic concatenation procedure called simultaneous probabilistic conjoint additivity to construct an empirical dimension that conforms to fundamental measurement axioms. A separation of items and observations during this process establishes a measurement framework with properties of invariance and objectivity. Rasch model transformation is mathematically based on $\beta$ and $\delta$ in the following expression,

$$
\Pi_{nix} = \frac{\exp \sum [\beta_h - (\delta_i + \tau_j)]}{\sum \exp \sum [\beta_h - (\delta_i + \tau_j)]}
$$

where $\beta$ = observations, $\delta$ = item difficulties, and $\tau$ = rating scale thresholds. $\Pi_{nix}$ is the probability any item $\delta_i$ will be rated X by participant $\beta_h$, where X takes a value from a fixed range ($j = 1, 2, 3, 4, 5, 6, \ldots n$), m = number of steps for an item, and k = ith step. Model prediction (P) for each item and observed ratings (O) are statistically analyzed (O-P) for significant departures from expectation with Chi-square analysis. When raw data fit a Rasch model, differences between observations have an explicit unit of measurement with axiomatic additivity independent of reference samples. Statistical estimation of model parameters implements an empirical probabilistic concatenation procedure.
4.2. Measurement of change
Raw difference scores were calculated where $Y_1 = \text{pretest scores}$, $Y_2 = \text{posttest scores}$, and $D = \text{difference}$:

$$D = Y_2 - Y_1$$

Finally, raw scores were converted to residualized gains with a general model:

$$Z_{y,x} = Z_y - r_{xy}Z_x$$

Where $Z_y$ represents post-test scores in z-score form, $r_{xy}$ is correlation between pre- and post-tests, and $Z_x$ represents pre-test score in z-score form. This model was implemented with regression procedures, where gain is deviation of Time 2 from regression line after regression of Time 2 on Time 1.

In this study, a residualized gain score is defined by $Y - Y'$, where $Y'$ is regression of $Y$ on $X$:

$$Y' = E(Y) + \frac{\text{Cov} (X, Y)(X - E(X))}{\text{Var} X}$$

where $E$ denotes the expectation, $\text{Var}$ the variance, and $\text{Cov}$ the covariance.

4.3. Sample
Observations were collected of several hundred children (4 year olds, mixed gender, and multi-racial) attending neighborhood preschools in Chicago, USA.

4.4. Data
School performance of children was observed with CC.NET, which is a commercially prepared early childhood assessment with 50 rating scale items [20]. Preschool Mini-Assessment (PMA) is a standard 35 item child interview protocol developed by Office of Early Childhood Education.

4.5. Procedure
Fall and spring (Time 1/Time 2) CC.NET and PMA observations were conducted with separate samples. CC.NET was collected by preschool personnel, while PMA interviews were conducted by evaluators.

4.6. Analysis
All ratings were summed to establish rank order, and then transformed to objective, linear scale values with a Rasch model for rating scales [21] that was implemented with WINSTEPS software (22). Then measurement of change was calculated by three methods, a) raw score gains, b) residualized raw score gains, and c) objective linear differences statistically controlling for pre-test differences.

5. Results
Figures 1 and 2 compare raw score and linear gains.

5.1. Raw score and linear gain
Ordinal methods assume uniform units for both easy and difficult items. Figure 1 shows CC.NET ratings crowded together near ceiling, which suggests this sample was able and items were “easy”. In contrast, linear measures show many children were lower ability and far fewer children were higher ability. Distance between items near ceiling is worth several linear units compared to ratings. Linear measures also show many children in the fall near CC.NET mid-range barely advanced any higher in the spring. In general, ratings show substantially more growth than linear measures. Both ratings and linear measures show some children “leap” from very low fall scores to spring ceiling. However, linear measures appear to
“push back” on fall/spring raw score differences, which suggest children are passing easy items but their ability is not advancing. Depending on locations in fall distribution, CC.NET gains based on linear measures differ significantly from ratings. Significant differences also appeared when PMA raw scores were compared with linear PMA measures. PMA results showed most children advanced uniformly across the ability distribution between fall and spring, while linear measures indicated a much smaller advance restricted mainly to lower scoring children. Higher PMA scoring children in fall remained high in spring but did not show substantial growth. Linear measures show large gains for some low scoring children but many children do not show any advance between fall and spring. In other words, raw scores tend to "puff up" gains for everyone, while linear measures distinguish between passing easy versus difficult items. These differences between raw scores and linear measures have significant implications for program evaluation. In Figure 2, Plot A, PMA logit differences (Time 2-Time 1) and standardized residualized gain scores are remarkably similar and consistent with simple linear transformation. Plot B shows little difference between PMA residual gains (Y-Y') and raw score differences \( Y_2 - Y_1 \). Plot C, however, shows little coherence between CC.NET residualized gain scores (Time 2-Time 1) and linear differences, while Plot D shows residualized gains (Y-Y') and raw score differences \( Y_2 - Y_1 \) very similar.

6. Discussion
These results challenge simple raw score comparisons and measurement of change. Results show remarkable consistency between raw scores and linear measures when initial raw scores are concentrated in a relatively narrow range. This consistency was especially apparent when PMA Time 2 raw scores were concentrated in lower mid-range and gain magnitude was generally modest. When raw score gain was more broadly distributed such as CC.NET, measurement of change degraded profoundly and relationship between raw and linear gain was absolutely incoherent. Residualized transformation did not offer significant improvement to raw score measurement of change for either PMA or CC.NET. Ontological benefits of linear measures in this research consist of parameterized hierarchical item structures, which should be useful for advancing early childhood learning theories.

7. Conclusions
A general assumption of this paper is numerals assigned to observations by fundamentally different mathematical expectations such as uniform raw score counts and probabilistic Rasch models should yield measures with profoundly different epistemological consequences. In fact, based on comparisons described above, precision, simplicity, ontology, and meaningfulness were found to differ significantly between raw scores and linear measures. These results are summarized below.

7.1. Meaningfulness of measuring change. Meaning, difficulty, and validity of raw scores are unstable over time and conditions, which distort measurement of change. Amount of distortion is related directly to initial status and magnitude of gain. Objectivity is compromised because slight differences in sample locations profoundly affect raw score gain. Methods that cannot account for initial status should be avoided. Residualized gain adaptations of raw scores did not offer significant advantages over simple raw scores. Measurement of change at group and child level is incommensurable.

Raw score assessments with substantial ceiling effects at post test may show "gain inflation", while assessments with central distributions show less distortion. This issue, however, is problematic because raw scores after successful interventions will always shift and items become easier, which compromises invariance of the measurement framework. Under these conditions, valid measurement requires explicit item equating methods to stabilize measurement frameworks between measuring occasions.

Validity issues related to non-uniform units seriously limit usefulness of untransformed difference scores and residualized gains. Empirical analysis here found them statistically identical. Moreover, these results suggest all untransformed raw score derivatives should be avoided. Differences between ordinal scores and linear measures have practical implications for pre-post program evaluation, standard setting methods, and clinical standards to address minimally important differences.
Figure 1. Time 1 and Time 2 Comparison of CC.NET Ratings and Linear Measures.
Figure 2. Comparison of Linear Measures (Rasch logits) and Residualized Gains.
7.3. **Conceptual entities.** Probabilistic measurement models in social research commonly assert latent traits, which are operationally defined by mental ability and attitude questionnaire items. When empirical observations conform to linear model expectations, these latent traits acquire status as conceptual entities. Their empirical reality becomes observable when statistical parameter estimation manifests a calibrated item-observation map. In this research, linear model fit statistics indicated Rasch model dimensions have ontological status that is substantially more coherent than raw score results. Based on these results, linear models assert heuristic properties that extend and elaborate scientific knowledge about learning and development of young children.

7.4. **Intra-subject change.** Weak measures depend on aggregation to describe sample variability, which sacrifices measurement invariance and by implication objectivity. Moreover, aggregation confounds sample variability with intra-subject change across multiple measurements. Raw score failure to maintain uniqueness systematically violates measurement integrity. In contrast, objective, linear measures show superior capacity to describe intra-subject change independent of items/samples.

7.5. **More stable and less restrictive knowledge.** Measures based on invariant parameter estimation in linear probability models are less conditional and restricted than raw scores. They are fundamental to meaningful knowledge structures in complex psychosocial systems.

**Acknowledgements**

The commitment of Office of Early Childhood Programs, and specifically, Barbara Bowman, Chief Officer, Paula Cottone, Deputy Chief Officer, and Karen Carradine, Director of Programs to research and evaluation made this study possible. We are deeply grateful for their permission to present these results.

**References**

[1] Wright B D and Linacre J M 1989 *Arch. Phys. Med. Rehab.* 70 857-60
[2] Fan X 1998 *Educ. Psychol. Meas.* 58 357-81
[3] Richters J E 1997 *Dev. Psychopathol.* 9 193–229
[4] Luce R D and Tukey J W 1964 *J. Math. Psychol.* 1 1-27
[5] Barrett P 2008 *Measurement* 6 78–123
[6] Andrich D 2004 *Med. Care* 42 1-7 to I-16
[7] Rossi G B 2006 *Measurement* 39 512–21
[8] Borsboom D and Mellenbergh G J 2004 *Theor. Psychol.* 14 105–20
[9] Borsboom D Mellenbergh G J and Heerden J V 2004 *Psychol. Rev.* 111 1061–71
[10] Finkelstein L 2009 *Measurement* 42 1270–77
[11] Wright B D 1999 *New Rules of Measurement* ed S E Embretson and S L Hershberger (Mahwah, N J: Erlbaum)
[12] Molenaar P C M 2008 *Dev. Psychobiol.* 50 60–9
[13] Hobart J C Cano S J Zajicek J P and Thompson A J 2007 *Lancet Neurol.* 6 1094–105
[14] Williams R H and Zimmerman D W 1996 *Appl. Psych. Meas.* 20 59-69
[15] Mellenbergh G J 1999 *Appl. Psych. Meas.* 23 87-9
[16] Zimmerman D W and Williams R H 1982 *J. Educ. Meas.* 19 149-54
[17] Rogosa D R and Willett J B 1983 *J. Educ. Meas.* 20 335-43
[18] Kissane, B V 1982 *Educ. Res. Perspect.* 9 55-72
[19] Rasch G 1960/1980 *Probabilistic Models for Some Intelligence and Attainment Tests* (Chicago: Univ. of Chicago Press)
[20] Dodge D T Colker L J and Heroman C 2001 *Developmental Continuum Assessment System* (Washington, DC: Teaching Strategies)
[21] Wright B D and Masters G N 1982 *Rating Scale Analysis* (Chicago: MESA Press)
[22] Linacre J M 2005 *WINSTEPS: Rasch Model Software and Manual* (Chicago: Winsteps.com)