

Borrowing from Peter to Score Paul: Issues in Measurement, Score, Reliability and Replication

Megan Cox
University of Minnesota

Katherine McCormick
University of Kentucky

Beth Rous
University of Kentucky

Purpose: To examine the appropriateness of replication of published measures for children with disabilities and their families. Attend to the need for increased operationalization of quality indicators in measurement.

Objectives:

1. Investigate differences in reliability and validity studies across suggested quality indicators (i.e., Division of Research Task Force on Quality Indicators for Special Education; Odom, Brantlinger, Gersten, Horner, Thompson, & Harris, 2004).
2. Determine need for guidance to enhance reporting of reliability statistics across populations and studies to increase the evidence base and utility of established measures in early childhood.

Comparison of reliability across populations:

To expand instrument utilization across populations for comparable measurement of outcomes, it is important that we report and replicate score reliability coefficients from studies involving multiple populations.

Method: The data for this study is a subsample of a larger data set of a multi-state study conducted by the National Early Childhood Transition Center. Typically data were collected in the child and family's home by trained data collectors. For more details about NECTC, go to:

<http://www.hdi.uky.edu/NECTC/Home.aspx>

Measures*:

- Family Support Scale (Dunst, Jenkins, & Trivette, 1984).
- Family Empowerment Scale (Koren, DeChillo & Friesen, 1992)
- Early Intervention Services Assessment (Aytech, Castro & Selz-Campbell, 2004)

All measures provided evidence of construct validity and reliability within their study populations.

Little attention has been paid in published literature to cross population reliability (Vacha-Haase, Henson & Caruso, 2002). This is of greater concern when a measure is valid, but reliability for different populations has not been established.

*Measures were chosen based on the existence of evidence that previous factor analysis and internal consistency had been completed.

Discussion Question 1: How much evidence is enough to call a measure evidence-based?

The following trends emerged with regard to evidence on measures as the result of a research synthesis on early childhood transition completed by NECTC:

Of 18 family based transition studies identified and reviewed :

- 10 used project developed measures but did not report reliability or validity data;
- 8 used published instruments. Only three of these 8 provided sample-specific reliability data on the instruments used.

NECTC Sample: 179 children and families exiting Early Intervention Services:

- Child Gender: 63.6% male
- Child Race: 33% non-white
- Average age in months: 30.96
- Most common diagnosis: 32% Speech & 28% DD
- Respondents: 81% biological parents

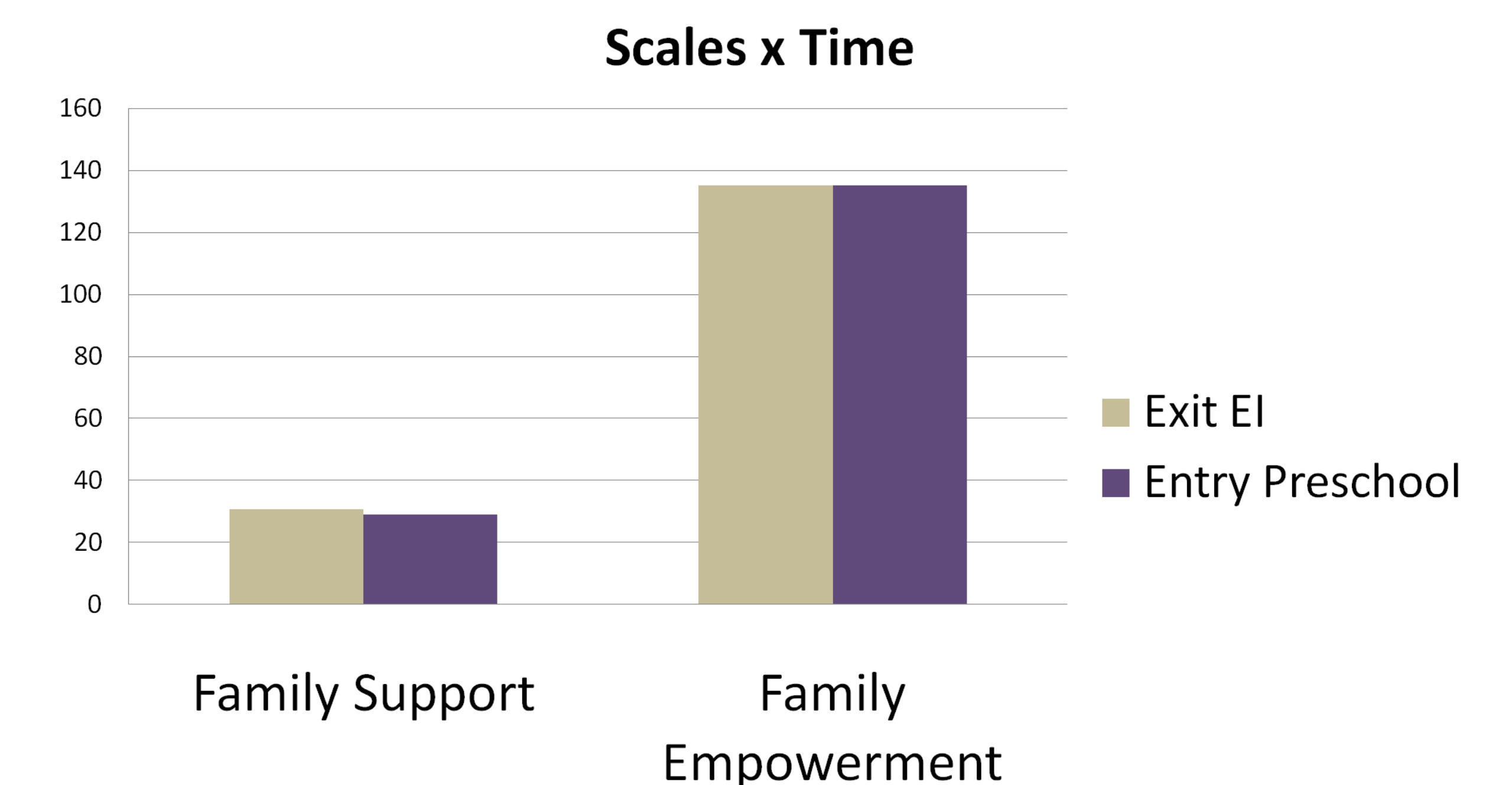
Reliability:

	Original Sample	Factors	Other Samples	NECTC sample	Notes
Family Empowerment Scale**	.87-.88	3	.93	.94	NECTC & Other samples yielded different factor loadings
Family Support Scale**	.77	5	.85	.93	NECTC factors were consistent with previous findings
EISAS Transition Subscale	n/a	1	n/a	.87	NECTC samples yielded different factor loadings

**Measures had established evidence with satisfactory factor loadings and internal consistency prior to NECTC study

Discussion Question 2: How similar is similar? How many variables should be considered when investigating the comparability of samples?

The previous scales were investigated across sample characteristics for reliability and factor loadings. Each sub-sample had similar internal consistency and factor loadings for each measure. In addition no significant differences occurred over time within the Family Support Scale and Family Empowerment Scale.



Discussion Question 3: Is there a time when score reliability can be assumed for a set of scores/items? How much is enough? Who determines this? When does the field determine that score reliability can be assumed for a set of scores/items? Is this different for types of instruments – Norm referenced, curriculum-based checklists,? For example – the PPVT enjoys widespread use – is score reliability assumed for this instrument?

Discussion Question 4: What steps can be taken to more closely approximate a replication of score dispersion from previous studies? Why is this important?