Linking Measure Scores from the FOTO Neck Functional Status Computer Adaptive Test (Neck CAT) and the Neck Disability Index (NDI)

An illustration of research methods and results behind the FOTO crosswalks

Michael A. Kallen, PhD, MPH
Psychometric Research Scientist, Net Health Systems, Inc., Pittsburg, PA
Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

June 2018

INTRODUCTION

Linking establishes a functional statistical relationship between two measure score distributions and their score scales. When measures are not created according to the exact same test specifications (e.g., they differ in length or item content), their linked scores are similar but not interchangeable. However, the linked scores are statistically related to one another, and the scores from one measure can be predicted from scores of the other. Linked scores provide robust, group-level summary information. The purpose of the analyses reported here was to link the scores of the FOTO Neck CAT and scores of the Neck Disability Index (NDI). Below we describe important properties of successful (robust) linking, and we evaluate our results regarding these properties.

LINKING FINDINGS

Linking Samples

We used two samples for our linking investigations. Each sample took both the FOTO Neck CAT and the NDI. This allowed us compare how patients actually scored on a measure to how they were predicted to score by the linking results. The first sample (N=13,792) was used to conduct the initial linking analyses. The second sample (N=1000) was used as a Validation Sample. In the Validation Sample, we obtained linked scores in this independent sample and then compared the characteristics of those scores to the characteristics of the linked scores from the original Linking Sample. Similar linked score characteristics across independent samples is an indication of successful linking.

Reliability and Standard Error of Measurement (SEM)

Measure scores are not perfect estimates; there is always some amount of error associated with scores. It is not surprising, then, that scores from each linked measure have their own reliability characteristics, demonstrating good (hopefully!) but not perfect reliability. The reliability of scales impacts how precisely they measure. Therefore, it is helpful to report both reliability estimates, such as Cronbach's alpha, and standard errors of measurement (SEMs). SEMs are reported in the units or score points of the linked measure, and thus are a measure-specific indication of score precision. Another point to be aware of is that scoring error "accumulates" when two measures are linked. Linking scores have error associated with each of the individual measures plus linking error. Table 1 reports the results from linking in the original (Linking) sample and in the Validation Sample.

Table 1: Reliability and Standard Error of Measurement (SEM) of the Linked Measures

NECK		Neck CAT				NDI		
Sample	**N**	**SD**	**Median SE**	**Reliability** *	**SEM**	**SD**	**Reliability** +	**SEM**
Linking	13792	12.351	3.736	0.909	3.736	16.939	0.855	6.450
Validation	1000	12.257	3.736	0.907	3.736	17.119	0.860	6.405

*Standard Error (SE)-based reliability estimate

+Cronbach's alpha reliability estimate

Measure Score Distributions

We used a linking methodology that is robust to the "shapes" of the measure score distributions to be linked. The method works well whether score distributions are normal, skewed, or have excessive kurtosis. Nevertheless, it is useful to evaluate the shapes of the score distributions to be linked and to confirm that the predicted (linked) score distribution displays characteristics similar to those of the related actual (observed) score distribution. Below are the observed score distributions for the FOTO Neck CAT (Figure 1) and for the NDI (Figure 2).

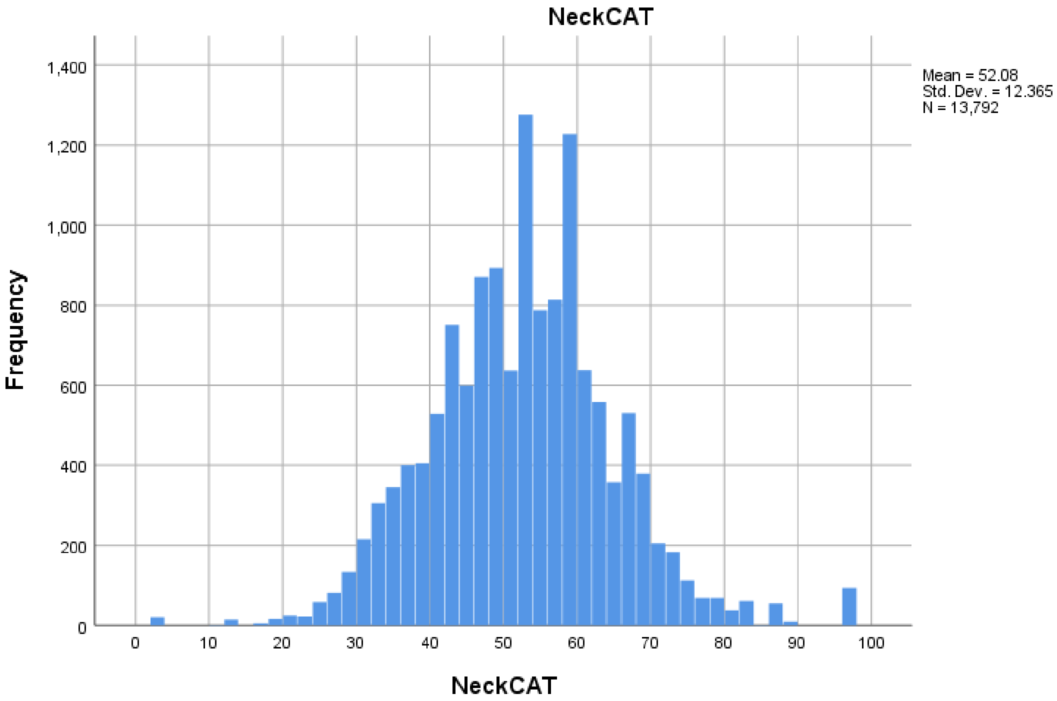


Figure 1: Score Distribution of the Neck CAT Measure (Linking Sample)

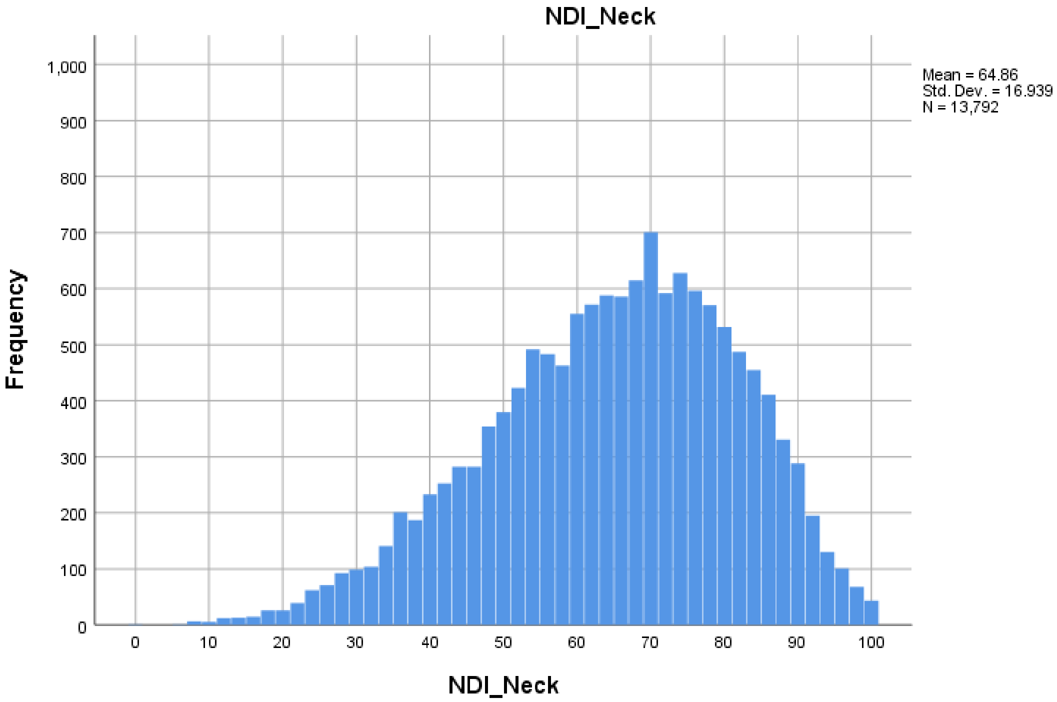


Figure 2: Score Distribution of the NDI Measure (Linking Sample)

We expect that the characteristics in the two “source” distributions will be expressed in the distribution of the linked scores. Figure 3 shows that, in the current study, this was the case. Visually, the predicted (linked) NDI score distribution is highly similar to the actual (observed) NDI score distribution, while it retains some of the score-frequency characteristics of the source FOTO Neck CAT score distribution.

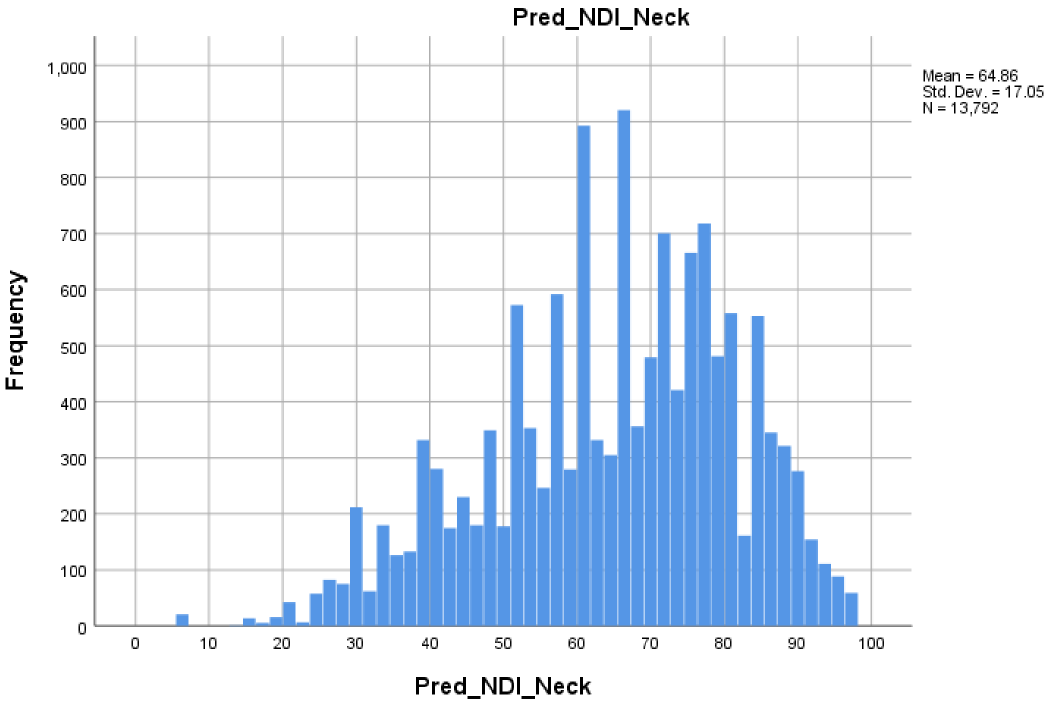
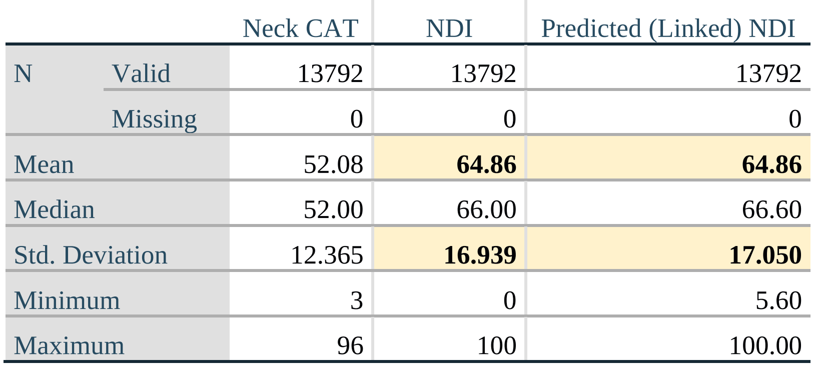


Figure 3: Score Distribution of the Predicted or Linked NDI Measure (Linking Sample)

Table 2 reports each score distribution’s descriptive statistics. Here we also see evidence of the high similarity between the predicted (linked) NDI score distribution and the actual (observed) NDI score distribution.

Table 2: Distribution Characteristics of the Linked Measures (Linking Sample)



		Neck CAT	NDI	Predicted (Linked) NDI
N	Valid	13792	13792	13792
	Missing	0	0	0
Mean		52.08	**64.86**	**64.86**
Median		52.00	66.00	66.60
Std. Deviation		12.365	**16.939**	**17.050**
Minimum		3	0	5.60
Maximum		96	100	100.00

Score Ranges

Another property we hope to see is that the range of linked scores includes the full range of possible scores for the measure being linked. This is an indication of successful linking and, happily, we found this property nearly perfectly expressed in the current study. NDI scores range from 0-100, while the predicted (linked) NDI scores range from 5-100 (Table 2).

Measurement of Equivalent Constructs

A fundamental requirement for linking measure scores is that the two measures to be linked assess essentially equivalent constructs. The equivalence of what two scales measure can be evaluated by correlating scores from the two measures. Ideally, we want the scores to be correlated around 0.80 or higher; however, measure score correlations between 0.60 and 0.79 may form the basis of constructive linking. We also have expectations about how well linked scores correlated with actual scores. Our expectations for this correlation is that they will be as high as the correlation between observed scores on the two measures or slightly higher. Table 3 shows that this is what we found in linking the FOTO Neck CAT and the NDI.

Table 3: Pearson Correlations between the Linked Measures (Linking Sample)

		Neck CAT	NDI	Predicted (Linked) NDI
Neck CAT	Pearson Correlation	1		
	Sig. (2-tailed)			
	N	13792		
NDI	Pearson Correlation	.684**	1	
	Sig. (2-tailed)	.000		
	N	13792	13792	
Predicted (Linked) NDI	Pearson Correlation	.982**	.694**	1
	Sig. (2-tailed)	.000	.000	
	N	13792	13792	13792

**. Correlation is significant at the 0.01 level (2-tailed).

VALIDATION SAMPLE

In the second part of our study, we repeated the analyses conducted with the original Linking Sample. Obtaining similar results would suggest the generalizability of our results. That is, it would indicate that our findings are not just good in one sample.

Measure Score Distributions

As with our Linking Sample, the linked score distributions in the Validation Sample displayed characteristics similar to those of the FOTO Neck CAT and the NDI score distributions. Figures 4-6 show the results. Visually, the predicted (linked) NDI score distribution is quite similar to the actual (observed)

NDI score distribution, while it has also retained some of the score-frequency characteristics of the source FOTO Neck CAT score distribution.

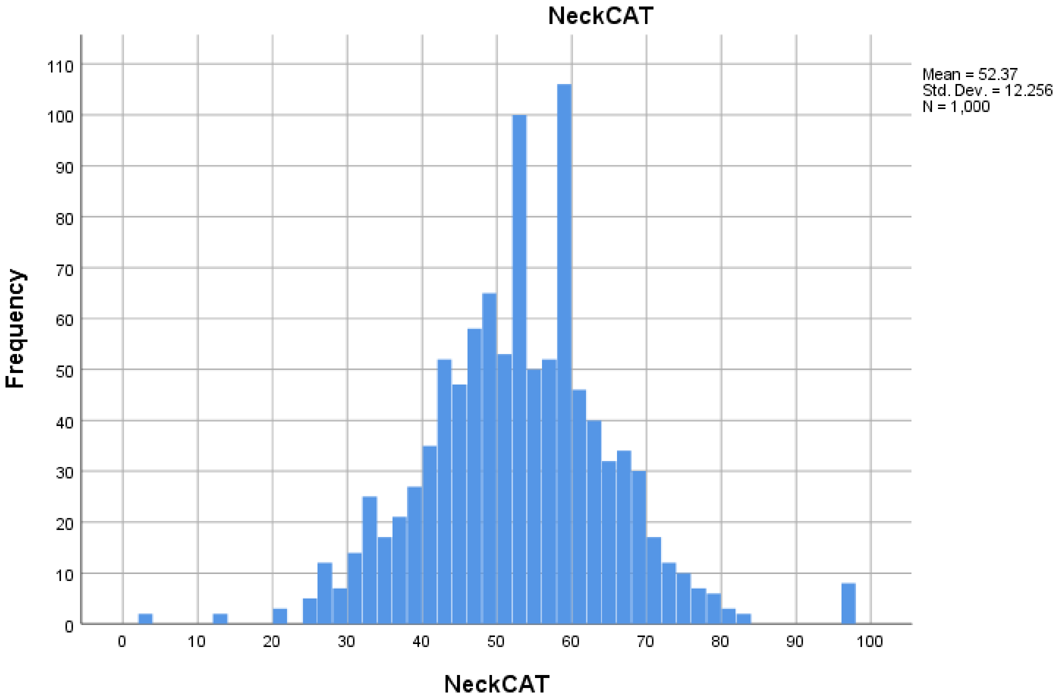


Figure 4: Score Distribution of the FOTO Neck CAT Measure (Validation Sample)

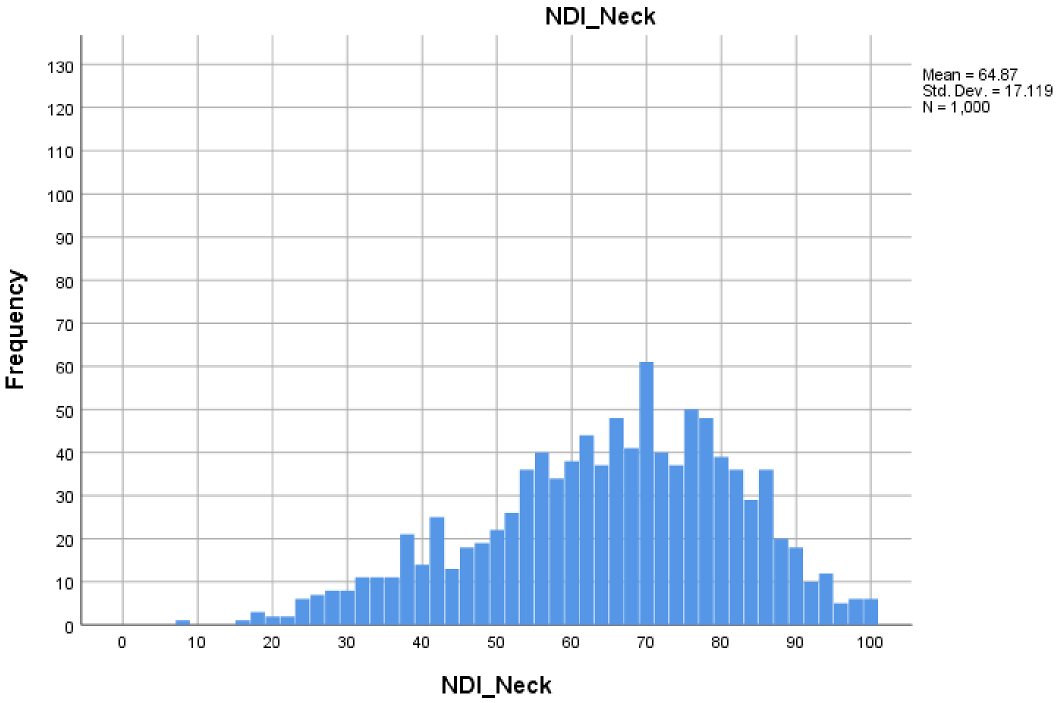


Figure 5: Score Distribution of the NDI Measure (Validation Sample)

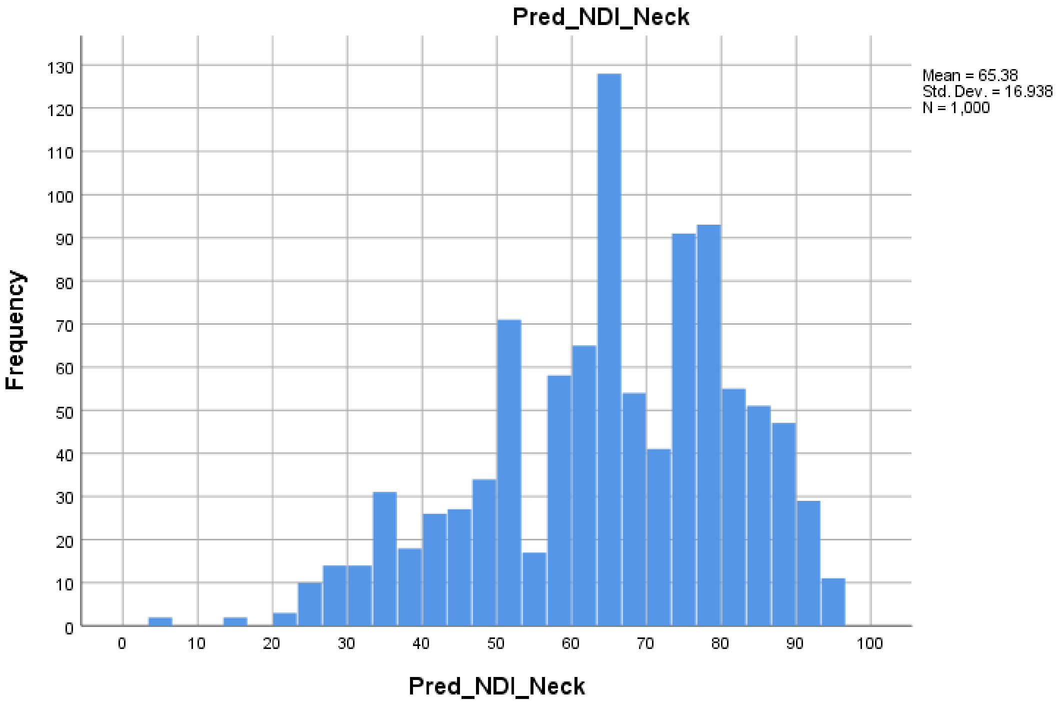


Figure 6: Score Distribution of the Predicted or Linked NDI Measure (Validation Sample)

In Table 4, we see that the descriptive statistics for the Validation Sample's predicted (linked) NDI scores and actual (observed) NDI scores are quite similar to those of the Linking Sample's predicted and actual

NDI scores (compare with Table 2). This further supports the robustness and generalizability of our linking: that, across samples, linked scores have similar characteristics to observed scores.

Table 4: Distribution Characteristics of the Linked Measures (Validation Sample)

		Neck CAT	NDI	Predicted (Linked) NDI
N	Valid	1000	1000	1000
	Missing	0	0	0
Mean		52.37	**64.87**	**65.38**
Median		52.00	66.00	66.60
Std. Deviation		12.256	**17.119**	**16.938**
Minimum		3	8	5.60
Maximum		96	100	100.00

Score Ranges

As Table 4 shows, similar to the linking sample, the validation sample had the desirable property of exhibiting the full possible score range (0-100) in the linked scores (5 to 100).

Measurement of Equivalent Constructs

We observed that scores on the FOTO Neck CAT had a correlation of 0.681 with observed NDI scores, similar to the Linking Sample's Neck CAT-NDI correlation of 0.684. We expected and observed that Predicted (Linked) NDI scores correlated with actual (observed) NDI scores at or slightly above the magnitude of the correlation obtained in support of linked-measure construct equivalence (i.e., 0.681).

Table 5: Pearson Correlations between the Linked Measures (Validation Sample)

		Neck CAT	NDI	Predicted (Linked) NDI
Neck CAT	Pearson Correlation	1		
	Sig. (2-tailed)			
	N	1000		
NDI	Pearson Correlation	**.681****	1	
	Sig. (2-tailed)	.000		
	N	1000	1000	
Predicted (Linked) NDI	Pearson Correlation	**.982****	**.690****	1
	Sig. (2-tailed)	.000	.000	
	N	1000	1000	1000

**. Correlation is significant at the 0.01 level (2-tailed).

Predicted (Linked) vs. Actual (Observed) NDI Individual Score Differences

Finally, we ask a most relevant question: "How similar are Predicted (Linked) and Actual (Observed) NDI scores?" Descriptive statistics reported in Table 6 help us to quantify this. "Mean Difference" is the average score difference, across the full score distribution, between actual and linked scores. The "Standard Deviation of the Difference" (SD Difference) estimates how variable the individual score differences were. The "Root Mean Squared Difference" (RMSD) is the square root of the average of squared errors (i.e., squared score differences); in other words, the RMSD is the average error of individual predicted (linked) scores, reported in the units (score-point values) of the NDI measure.

Another helpful statistic is Krippendorff's alpha reliability. This statistic estimates reliability based on "agreement" between scores (here, predicted vs. actual NDI scores). Finally, the Limits of Agreement (LOA) values form a confidence interval around expected score differences. The interval is based on the SD Difference (i.e., SD Difference X 1.96) and is centered on the Mean Difference.

Table 6: Difference Characteristics of Predicted (Linked) NDI Scores vs. Actual (Observed) NDI Scores

NECK		**Predicted (Linked) NDI *minus* Actual (Observed) NDI**			
Sample	**N**	**Mean Difference**	**SD Difference**	**Mean Squared Difference**	**Root Mean Squared Difference**
Linking	13792	0.001	13.295	176.756	13.295
Validation	1000	0.519	13.410	179.918	13.413
		Krippendorff alpha Reliability	**Limits of Agreement (+/-)**	**+ Boundary: Limits of Agreement**	**- Boundary: Limits of Agreement**
Linking	13792	0.694	26.058	26.059	-26.057
Validation	1000	0.690	26.284	26.803	-25.765

In Figure 7 and 8, the LOAs from the Linking and Validation Samples, respectively, are used to define Bland-Altman plots. These plots visually depict observed score differences, with upper and lower LOAs defining a "zone" of expected or acceptable score differences. We use these plots to diagnose the linking results at the individual-score level. For scores exceeding the LOA: How many such score differences occur? In what range of the measurement continuum do they occur?

For the Linking Sample, the large majority of score differences lie within the LOA zone. The ones that do not congregate mostly in the mid-section of the scoring continuum, not at its endpoints.

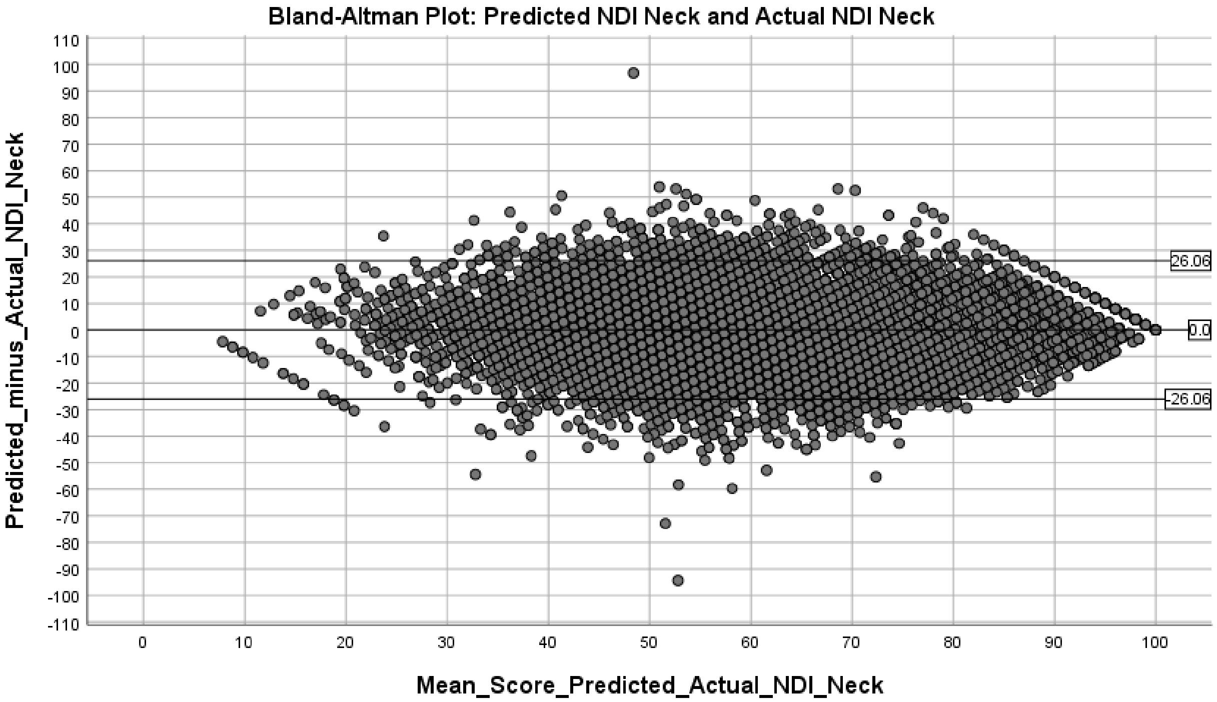


Figure 7: Bland-Altman Plot of the Differences between Predicted vs. Actual NDI Scores (Linking Sample)

For the Validation Sample (Figure 8), we see a very similar plot to the one developed using the Linking Sample results. The majority of Validation Sample score differences lie within the LOA zone, and those outside it are closer to the mid-section of the scoring continuum.

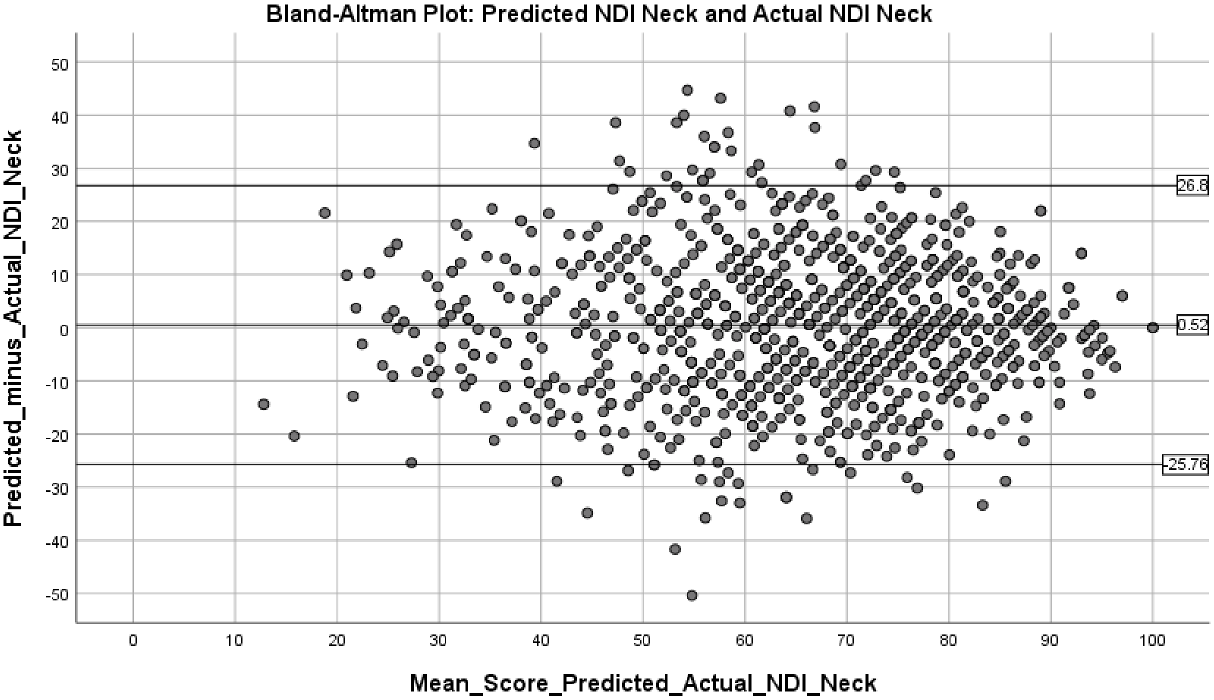


Figure 8: Bland-Altman Plot of the Differences between Predicted vs. Actual NDI Scores (Validation Sample)

Summary and Conclusions

The results from this study confirm that we have robust links between scores on the FOTO Neck CAT and scores on the NDI. The measurement properties we examined (similarity of score ranges and distributions, correlations among scores, and differences between actual and linked scores) all confirm successful linking. As Figures 8 and 9 show, however, there is substantial variability at the individual-score level. This underscores our recommendation that linked scores be used for sample comparisons, not for individual comparisons. Therefore, it would be appropriate, for example, to compute the linked score for each patient in an overall sample of 25 or 50 patients, but results should only be reported and used at a robust subgroup or overall sample level.