

3. VALIDITY AND RELIABILITY OF THE ESI SCALE AND THE ESI QUALITY OF SOCIAL INTERACTION MEASURES

3.1 Current Standardization Sample

The sample used to standardize the current version of the ESI was comprised of 1140 persons. In order to be able to describe the standardization sample in a meaningful manner, we have grouped the sample into six global groups, as follows:

1. ***Well*** — persons who are typically-developing, with no known or suspected disabilities or medical diagnoses, and who have no known medical needs or environmental factors that place them at risk for developmental delays or disabilities.
2. ***Children at risk for or persons with mild disabilities*** — children who are “at risk” are those who do not have any identified disability, or medical or educational diagnoses, but who (a) have medical needs or environmental factors that place them at risk for developmental delays, or (b) are experiencing learning, behavioral, or social problems in general education classrooms (Heward, 2006; Johnson, 1998). Those who have mild disabilities are persons who have identified disorders of attention (ADHD), specific learning disabilities ***not*** associated with mental retardation (i.e., intellectual disability) (LD) (e.g., dyslexia, speech and language disorders, reading disabilities), developmental coordination disorders (DCD), and/or sensory integrative disorders (SI). The two groups have been combined because the two groups are “virtually identical” in terms of their behavioral characteristics (Johnson, p. 224).
3. ***Psychiatric disorders*** — persons who have psychiatric disorders other than ADHD (e.g., schizophrenia, anxiety disorders, bipolar disorder, major depression, other affective disorders, personality disorders). This group also includes those with autism spectrum disorders.

4. **Developmental disabilities** — persons with identified developmental disorders including mental retardation, cerebral palsy, spina bifida, or other multiple or unspecified developmental disabilities.
5. **Neurological disorders** — persons who have acquired neurological disorders, including strokes and traumatic brain injuries.
6. **Other/multiple** — persons with (a) medical conditions; (b) dementia; (c) multiple disorders (e.g., cerebral palsy with visual or auditory impairments, mental retardation with concurrent musculoskeletal disorders, rheumatoid arthritis with or without concurrent neurological disorders); or (d) diagnoses that were unknown.

Approximately two-thirds of the standardization sample was comprised of healthy, well persons 2 to 90 years of age. Almost half of the sample (45.5%) was male; and 53.0% were tested in Nordic countries, 18.9% were tested in the United Kingdom and the Republic of Ireland, 9.6% were tested in Australia and New Zealand, 9.6% were tested in Asian countries, and 8.9% were tested in North America (United States and Canada). The demographic characteristics of the sample are summarized in Tables 3–1 and 3–2.

3.2 Evidence of Equivalence Between Gender Groups

When an instrument is used to test both males and females, it is important to verify that the normative values and any interpretations of the results apply equally to both genders. In the following sections, we will summarize the research providing evidence for the equivalence of mean ESI quality of social interaction measures between genders as well as evidence that the ESI scale is free of gender-related differential item functioning (DIF).

3.2.1 Mean ESI quality of social interaction measures by gender

To determine if there were gender-related differences in ESI quality of social interaction measures among the well sample used to develop the current ESI normative values (see Section 3.3.2 and Table B–1, Appendix B), we implemented a 2 X 13,

Table 3-1 Age of the ESI Standardization Sample by “Diagnostic” Group

Age (years)	“Diagnostic” group						Total
	Well	Children at risk/mild	Psychiatric	Developmental	Neurological	Other/multiple	
2	9	0	0	0	0	0	9
3	26	4	3	2	0	1	36
4	38	13	0	1	0	0	52
5	48	9	0	1	0	0	58
6-7	50	10	6	5	0	3	74
8-10	56	6	7	3	0	3	75
11-15	30	3	4	6	0	2	45
16-21	21	1	18	6	4	9	59
22-29	80	1	27	7	6	6	127
30-39	125	1	42	4	11	9	192
40-49	85	0	41	5	27	10	168
50-59	79	0	15	1	29	14	138
60-64	29	0	3	0	4	14	50
65-99	27	1	4	0	6	19	57
Total	703	49	170	41	87	90	1140

Table 3-2 Gender and World Region of the ESI Standardization Sample by “Diagnostic” Group

	“Diagnostic” group						Total
	Well	Children at risk/mild	Psychiatric	Developmental	Neurological	Other/multiple	
Gender							
F	456	16	57	20	37	35	621
M	247	33	113	21	50	55	519
World region							
North America	39	11	13	22	6	10	101
United Kingdom and Republic of Ireland	129	0	57	3	2	24	215
Nordic countries	386	15	79	8	78	38	604
Australia and New Zealand	66	1	20	4	1	18	110
Asia	83	22	1	4	0	0	110
Total	703	49	170	41	87	90	1140

gender by age group, analysis of variance (ANOVA) for the well sample and found no significant age by gender interaction effect ($F[13,675] = 1.53, p = .10$) or significant main effect for gender ($F[1,675] = 1.58, p = .21$). These results indicate that there is ***no significant difference in ESI quality of social interaction measures between gender groups***, and that combined gender normative values shown in Table B–1, Appendix B can be used.

Further evidence of equivalence of ESI quality of social interaction measures between gender groups was provided by Søndergaard (2009), who found no significant gender effects ($F[1,90] = 1.43, p = .24$) among three gender-matched groups of adults who were well, or who had neurological or psychiatric disorders ($n = 32$ persons, 16 males and 16 females, in each group).

3.2.2 Differential item functioning by gender

When gender is considered, it is also important to verify that an instrument displays no evidence of ***differential item functioning (DIF) associated with gender***. DIF refers to the idea that some, but not all of the items or intended purposes of social interaction included in the ESI are relatively easier (or harder) for persons from one gender group than they are for persons from the other gender group. Conversely, when the item or intended purpose calibration values are the same among groups, those that are easy for one group are also easy for those in another group. While the presence of DIF is never ideal, it is not always severe enough to disrupt the measurement system. When it does, test bias occurs.

There is no universally accepted standard for determining if DIF is present. While Rasch computer programs often use t tests to evaluate for DIF, large sample sizes are associated with small standard errors and too much power, and the risk is over identification of significant differences. The more common alternative, therefore, is the use of effect sizes, but again, there is no commonly accepted standard. While recommended values typically range from 0.40 logit to 0.60 logit, it has become most common to consider values less than 0.50 logit as evidence of no differential item or task functioning (Conrad, Dennis, Bezruczko, Funk, & Riley, 2007; Linacre, 1994; Tennant & Pallant, 2007; Tristán, 2006; Wilson, 2005; Zwick, Thayer, & Lewis, 1999). Differences ≤ 0.43 logit are indicative of a negligible effect size (Wilson, 2005). See Section 2.2.2, Chapter 2 for a definition of the term *logit*.

To evaluate for gender-related DIF for the ESI items and intended purposes, we performed separate MFR analyses (Bond & Fox, 2007; Fisher, 1993, 1994; Linacre, 1993, 2009a, 2010b), one for items and one for intended purposes. The results of our analyses of the data for 621 males and 519 females in the ESI standardization sample revealed no ESI item difficulties or intended purpose challenges that differed by more than 0.50 logit between gender groups (maximum difference: items = 0.24 logit, intended purposes = 0.03 logit). We concluded, therefore, that there is *no evidence of gender-related DIF in the ESI*.

Current research indicates that the ESI is free of bias associated with gender:

- **There are no significant differences in mean ESI quality of performance measures between gender groups**
- **The ESI items and intended purposes display no evidence of gender-related DIF**

3.3 Evidence that Quality of Social Interaction Increases with Age

3.3.1 Correlation between quality of social interaction and age

Since quality of social interaction increases with age, one way to validate the ESI is to examine for evidence that ESI measures demonstrate an expected positive relationship with age. An initial evaluation of the relationship between ESI quality of social interaction measures and age among 30 well, typically-developing children, 2 to 8 years of age, revealed a significant positive relationship ($r = .91, p < .001$) (Johansson Jänkänpää & Malmhäll, 2008). A subsequent evaluation of the relationship between ESI quality of social interaction measures and age among 305 well, typically-developing persons, 2 to 78 years of age, revealed a *significant high positive curvilinear relationship with age* ($R^2 = 0.796, p < .001$) (Griswold, 2009). Finally, when the ESI quality of social interaction measures for the 703 well persons,

2 to 90 years of age, in the current standardization sample were analyzed, the results confirmed a significant curvilinear relationship with age ($R^2 = 0.767, p < .001$) (see Figure 3–1).

ESI quality of social interaction measures increase significantly with age until about age 15 years, plateau between about 20 and 60 years of age, and then possibly decline slightly with increasing age.

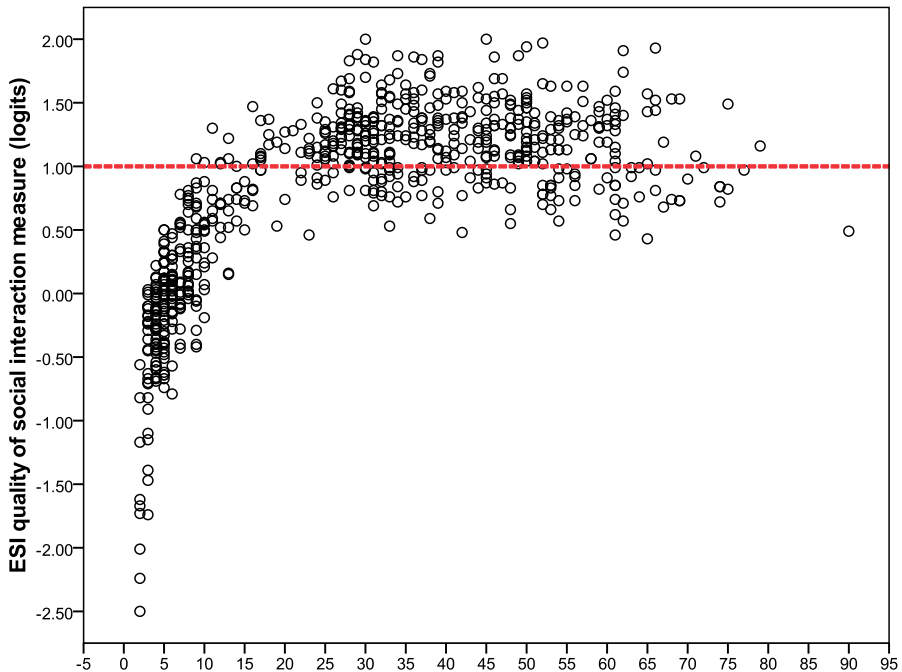


Figure 3–1. Scatterplot of the curvilinear relation between ESI quality of social interaction measures and age. The dashed line at 1.00 logit identifies the ESI cutoff for competent quality of social interaction (see Section 8.1.4, Chapter 8).

3.3.2 Mean quality of social interaction for well persons by age group

A complimentary method for examining the relationship with age is to compare the mean ESI quality of social interaction measures for well persons of different ages. The mean ESI quality of social interaction measures for the well persons in the ESI standardization sample, by age group, are shown in Table B–1, Appendix B. The results of an ANOVA analysis confirmed a significant relationship ($F[13, 689] = 1213.79, p < .001$). Post hoc Tukey *HSD* revealed that the mean ESI measure for 4-year-olds did not differ significantly from the mean ESI measure for 5-year-olds, and the mean ESI measure for 5-year-olds did not differ significantly from the mean ESI measure for 6–7-year-olds ($p \leq .05$). Otherwise, quality of social interaction increased significantly with age among children 2 to 15 years of age. In contrast, there were no significant differences mean ESI quality of social interaction measures among persons 16 years of age and above. It should be noted that some age groups have relatively small *ns*, which may mean that there was insufficient power to detect differences between age groups that may actually exist.

3.4 Evidence of Ability to Differentiate Among Groups

Another way to validate an instrument is to gather evidence that the measures generated by the instrument are *sensitive enough to differentiate among groups known to differ*. In the following sections, we will summarize evidence that ESI quality of social interaction measures differ among groups.

3.4.1 Well adults with different overall quality of social interaction

Even among healthy, well adults with no identified problems with social interaction, quality of social interaction varies. That is, well adults often display questionable or mildly inappropriate social interactions, and at times, even moderately inappropriate social interactions, and their ESI quality of social interaction measures often fall below the ESI cutoff for competent social interaction (see Figure 3–1). Among the 418 well adults in the standardization sample between 16 to 64 years of age, only 117 (28.0%) were judged by the occupational therapist to have appropriate (i.e., polite, respectful, well-timed, relevant, mature) social interactions, and to have caused no disruption or disturbance in their social interactions with their partners,

during all social exchanges observed (see Table 3–3). When the well sample was divided into four combined-gender groups (i.e., appropriate, questionable, mildly inappropriate, moderately inappropriate) based on each person’s *lowest* overall quality of social interaction, a one-way ANOVA revealed a significant main effect for overall quality of social interaction ($F[3, 414] = 66.89, p < .001$). Post-hoc Tukey *HSD* tests revealed that *all four groups of well adults differed significantly* ($p \leq .05$) in mean ESI quality of social interaction measures (see Table 3–3). These results support the sensitivity of the ESI quality of social interaction measures.

Table 3–3 Lowest Observed Overall Quality of Social Interaction and ESI Quality of Social Interaction Measures (logits) Among Well Adults 16–64 Years of Age

Overall quality	<i>n</i>	%	ESI quality of social interaction measure (logits)	
			<i>M</i>	<i>SD</i>
Appropriate, both tasks	117	28.0	1.43	0.26
Questionable, at least one task	149	35.6	1.23	0.24
Minimally inappropriate, at least one task	124	29.7	1.09	0.24
Moderately inappropriate, at least one task	28	6.7	0.82	0.14

3.4.2 Differences between well adults and adults with neurological or psychiatric disorders

Søndergaard (2009) implemented a pilot study comparing the ESI quality of social interaction measures among well adults and those with neurological or psychiatric disorders. She had three gender-matched groups, with 16 males and 16 females in each group (i.e., well, neurological, psychiatric). She found a significant main effect for group ($F[2, 90] = 31.27, p < .001$). Post-hoc Tukey *HSD* tests revealed that the groups with neurological or psychiatric disorders did not differ from each other ($p > .05$), but that both groups had significantly lower ESI quality of social interaction measures than did the well group ($p \leq .05$).

When we analyzed the ESI data for the well adults and those with neurological or psychiatric disorders, 16 to 64 years of age, in the current standardization sample, we again obtained a significant main effect for group ($F[2, 642] = 292.49, p < .001$). In contrast to Søndergaard's (2009) results, post-hoc Tukey *HSD* tests revealed that all three groups differed significantly in mean quality of social interaction ($p \leq .05$) (see Table 3–4). The group with psychiatric disorders varied more in overall quality of social interaction, and despite the fact that their mean age was lower than for those with neurological disorders, persons with psychiatric disorders had lower mean ESI quality of social interaction measures.

3.4.3 Differences between typically-developing children and children at risk or with mild disabilities

Finally, we compared the ESI quality of social interaction measures between well, typically-developing children and children in the at risk/mild group. The children ranged in age from 3 to 13 years ($M = 5.8, SD = 2.4$ years) and were matched for age, world region, and gender ($n = 30$ boys and 15 girls in each group). The results of an independent samples *t* test revealed that the *well children had significantly higher ESI quality of social interaction measures than did those in the at risk/mild group* ($t[88] = 2.39, p = .009$). The mean difference in ESI quality of social interaction measures between groups was 0.20 logit.

Current evidence supports the sensitivity of the ESI quality of social interaction measures with regard to detecting differences among groups:

- **Well adults vary significantly in quality of social interaction**
- **Well adults have overall significantly higher quality of social interaction than do groups of persons with disabilities that are likely to affect quality of social interaction (e.g., stroke, brain injury, schizophrenia)**
- **Typically-developing children have significantly higher quality of social interaction than do children at risk for or with mild disabilities**

Table 3–4 Lowest Observed Overall Quality of Social Interaction and ESI Quality of Social Interaction Measures (logits) Among Well Adults, and Adults with Neurological or Psychiatric Disorders

	Well	Neurological	Psychiatric
Overall quality			
Appropriate, both tasks	117	0	6
Questionable, at least one task	149	11	8
Minimally inappropriate, at least one task	124	30	33
Moderately inappropriate, at least one task	28	31	59
Markedly inappropriate, at least one task	0	9	40
ESI measure			
<i>M</i>	1.21	0.59	0.42
<i>SD</i>	0.29	0.40	0.53
Age			
<i>M</i>	39.6	44.6	35.8
<i>SD</i>	12.1	11.0	11.6

3.5 Evidence of Rater Reliability

While inter- and intrarater reliability cannot be separated from each other statistically within the MFR model of the ESI, they can be separated to some extent conceptually. That is, if all raters score each person they rate in a consistent manner, according to the scoring criteria in the ESI manual, they will demonstrate interrater reliability, and if a single rater scores consistently across all of the persons he/she scores, the rater will demonstrate intrarater reliability. Statistically, both interrater and intrarater reliability are indicated by goodness of fit statistics generated using MFR analyses (Bond & Fox, 2007; Fisher, 1993, 1994; Linacre, 1993, 2009a, 2010b).

More specifically, a rater will demonstrate goodness of fit when the scores he/she assigns to each ESI item are consistent with the expectations of the MFR models of the ESI, as follows:

- A rater is more likely to assign higher scores to easier social interaction items than to harder social interaction items,
- A rater is more likely to assign higher scores for intended purposes that involve less challenging social interactions than for intended purposes that involve more challenging social interactions, and
- A rater is more likely to assign higher scores to persons who have higher quality of social interaction than to persons who have lower quality of social interaction.

There were 98 raters who rated the 1140 persons in the ESI standardization sample. Among these raters, all but two raters (2%) demonstrated acceptable infit goodness of fit to the MFR model of the ESI ($MnSq \leq 1.4$ and $z < 2$). This means that ***98% of the 98 raters demonstrated inter- and intra-rater reliability when scoring the ESI, supporting high overall rater reliability.***

3.6 Evidence of Reliability of the ESI Measures

Reliability pertains to the stability of measures when testing procedures are repeated with a group of individuals or a single individual (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The two most common ways to evaluate reliability are the standard error of measurement and the reliability coefficient (Haertel, 2006). The ***standard error of measurement (SE)*** of each person's ESI measure is perhaps the most useful reliability index, as it provides an estimation of the potential within-person variation of a person's measure expressed in the same units that are used in the scale (e.g., logits) (Feldt & Brennan, 1989). More specifically, the *SE* "represents the standard deviation of a hypothetical set of repeated measurements on a single individual" (Feldt & Brennan, p. 105), such that there is a 95% chance that his/her measure falls within $\pm 2 SE$ from his/her obtained measure. Moreover, if two ESI measures differ by more than $2 SE$, the two measures can be said to differ significantly ($p \leq .05$) (Harvill, 1991). Finally, the smaller the *SE*, the more likely the generated measures will be reliable and sensitive indices of change (Bond & Fox, 2007).

The use of *reliability coefficients* is a more traditional method of reporting the stability of test scores (Haertel, 2006). Unlike the *SE*, which is based on a single test score, reliability coefficients are based on correlations between two replications of the evaluation procedures administered to the same group under different conditions (e.g., different occasions, different forms of the instrument). The most common methods for estimating reliability coefficients are test-retest, parallel forms, and split-half methods. While *test-retest reliability* pertains to the stability of test scores from the same test (same form) administered at two separate occasions, *parallel forms reliability* pertains to the stability of test scores obtained from two different forms of the same test. One of the most common methods for estimating the reliability coefficient for test scores of groups is based on estimating the reliability based on all possible split-halves of a set of test items in the form of *Cronbach's coefficient alpha* (Allen & Yen, 2002; Ary, Jacobs, & Razavieh, 2002; Crocker & Algina, 1986; Cronbach, 1951; Feldt & Brennan, 1989; Haertel, 2006).

When Rasch analysis computer programs are used, they generate reliability estimates in the form of a *Rasch equivalent of Cronbach's alpha* and an *SE* for each person's estimated measure. Rasch analysis computer programs also generate reliability estimates in the form of a *separation index*. The Rasch equivalent of Cronbach's alpha, commonly referred to as the *separation reliability coefficient*, is closely related to a separation index which reflects the replicability of the person placements along the measurement scale. The separation index can also be reported as *G*, which reflects how well a set of items spread the group of individuals tested into statistically distinct levels (strata) of ability when the strata are separated by 3 *SE* units. Finally, traditional estimates of Cronbach's alpha tend to be higher than the Rasch equivalent estimates, as the Rasch equivalent form does not take into account extreme scores that tend to inflate traditional estimates of reliability (Bond & Fox, 2007; Smith, 2001).

3.6.1 Parallel forms reliability

When the ESI is administered, the person is typically observed during at least two social exchanges that differ in some manner (e.g., intended purpose of the social interaction, social partner's overall quality of social interaction, familiarity of the social partner, age of the social partner, environment, time of day). Moreover, as discussed above, the person's overall quality of social interaction often varies between

the two observed social exchanges, even though they occur on the same day. Thus, by considering each social exchange to be a separate but parallel form of the ESI, we can use the results for each of two parallel social exchanges to estimate the reliability of the ESI quality of performance measures. Because the comparison of these two parallel forms reflects the influences of many different potential sources of error, *comparing the results for the first social exchange with the results of the second social exchange provides a conservative estimate of reliability of the ESI quality of social interaction measures* (Haertel, 2006).

When the ESI quality of social interaction measures of 455 persons based on the first social exchange observed were compared to the quality of social interaction measures based on the second social exchange, the *parallel forms reliability coefficient was $r = .86$* . Of those 455 persons, 93 persons (20.4%) were rated by the occupational therapist who observed their performances as having lower overall quality of social interaction during the second observation, 275 (60.4%) were rated as having the same overall quality of social interaction, and 87 (19.1%) were rated as having better overall quality of social interaction during the second observation. When only those who had the same overall quality of social interaction during each of the two observations were considered, the reliability coefficient increased somewhat to $r = .93$.

3.6.2 Rasch equivalent of Cronbach's alpha: Separation reliability

When the data for the current standardization sample were analyzed, the many-faceted Rasch equivalent of Cronbach's alpha was $R = .94$, again supporting very high reliability of the ESI quality of social interaction measures.

3.6.3 Separation index and G

The separation index which indicates how reliably persons in a sample would be hierarchically ordered if they were given a parallel form of the test (Bond & Fox, 2007). Again based on the current standardization sample, the person separation index for the ESI was 3.83, $G = 5.4$, indicating that the sample could be reliably divided into at least 5.4 distinct strata separated by 3 SE (Fisher, 1992; Smith, 2001).

3.6.4 Standard error of measurement (SE)

As noted above, the SE is often considered the most useful reliability index as it can be used to evaluate not only the reliability of a person's obtained ESI quality of

social interaction measure, but it also provides an index that reflects the sensitivity of the ESI measures in the context of evaluating change. The mean *SE* for those persons in the current standardization sample was 0.17 logit.

While the mean *SE* provides an overall estimate of the reliability of the individual ESI quality of social interaction measures, the magnitude of the *SE* varies along the length of a scale. The traditional *SE* (calculated indirectly from the reliability coefficient, *r*, for an entire sample) will be greatest near the center of a scale and gradually decrease toward the ends of the scale (Harvill, 1991). Rasch-based *SEs* (calculated directly for each person) vary such that the *SE* becomes larger at extreme upper and lower ends of the scale (Smith, 2001).

We, therefore, also calculated the mean *SE* for equal intervals along the ESI scale (see Table B–2, Appendix B). For those persons whose ESI quality of social interaction measures fall within the range delineated by grey shading, the overall standardization sample mean *SE* provides a reasonable estimate of the person’s actual *SE*; for those outside that range, larger *SE* values likely apply. For more information about how to use the *SEs* shown in Table B–2, Appendix B when interpreting the results of an *ESI Progress Report*, see Section 10–3, Chapter 10.

3.6.5 “Trade-off” between sensitivity and stability of the ESI measures

Test developers and test users alike need to carefully consider the “trade-off” between the sensitivity and the stability (e.g., parallel forms or test-retest reliability) of the results of an evaluation instrument. Tools that lack sensitivity are often not useful when used to evaluate change in performance after intervention; too much sensitivity can result in diminished reliability. Existing evidence suggest that the ESI quality of social interaction measures are both sensitive and stable.

The low *SE* for ESI quality of social interaction measures, combined with high reliability coefficients, supports the sensitivity of the ESI measures for detecting change, while retaining high stability between two different sets of results.

3.7 Evidence of Internal Scale Validity

When we use MFR analyses to evaluate the internal validity of the ESI quality of social interaction scale, we examine the extent to which the ESI social interaction items and intended purposes that comprise the scales demonstrate goodness of fit to the MFR model of the ESI (Bond & Fox, 2007; Fisher, 1993, 1994; Linacre, 1993, 2009a, 2010b). That is, as it is with raters, there are certain expectations that the data must meet. These expectations, as they pertain to the items and intended purposes included in the ESI, are that easier social interaction items and less challenging intended purposes are more likely to be scored higher than are harder social interaction items and more challenging intended purposes.

When we examined the results of the MFR analysis of the data for the ESI standardization sample, we found that 25 of the 27 ESI items and all six of the intended purposes demonstrated acceptable goodness of fit to the MFR model for the ESI ($MnSq \leq 1.4$, $z < 2.0$). The items *Discloses* and *Thanks* continued to misfit the many-faceted Rasch model of the ESI.

While the inclusion of misfitting items can be a threat to unidimensionality, principal components analysis (PCA) of the standardized residuals supported unidimensionality (Linacre, 2009b, 2010a). More specifically, 44.4% of the variance was explained by the Rasch factor, and the amount of variance explained by the first contrast did not exceed 25% of the variance explained by the items; the eigenvalue was = 2. Moreover, when we compared the ESI quality of social interaction measures for the 1140 persons in the standardization sample when they were based on all 27 items versus when they were based on only 25 items (with *Discloses* and *Thanks* omitted), the correlation between the paired measures was $r = .996$, and the mean difference between measures was 0.05 logit ($SD = 0.05$ logit). Finally, the measures for only 26 persons (2.3%) exceeded 0.17 logit, the mean *SE* for the standardization sample (see Section 3.6.4 and Table B-2, Appendix B).

Considered together, these results suggest that the inclusion of the two items that failed to demonstrate acceptable goodness of fit (*Discloses* and *Thanks*) did not disrupt the measurement system (i.e., did not result in differential test functioning) (Pae & Park, 2006). Thus, we concluded that the results support the internal scale validity of the ESI quality of social interaction scale, and indicate that the ***ESI can be used to evaluate a single, unidimensional construct — quality of social interaction.***

Nevertheless, we proceeded to examine the data to try to identify the reasons that *Discloses* and *Thanks* misfit. We found that the majority of the misfit was associated with unexpectedly low scores. That is, in most cases, persons who were expected to receive scores = 4 were given scores of 2 or 1. We therefore, have revised the scoring criteria for both items in an attempt to enhance reliability of scoring. Our plan is to continue to monitor (a) goodness of fit of these two items; (b) unidimensionality of the ESI items, as evaluated using PCA analyses; and (c) differential test functioning.

3.8 Evidence of Person Response Validity

In like manner, person response validity refers to the extent to which the persons who have been evaluated with the ESI demonstrate goodness of fit to the MFR model of the ESI (Bond & Fox, 2007; Fisher, 1993, 1994; Linacre, 1993, 2009a, 2010b). As with raters, intended purposes, and ESI items, persons will demonstrate goodness of fit if their patterns of response across all ESI items are such that:

- All persons are more likely to receive higher scores on easier ESI items and less challenging intended purposes than they are on harder ESI items and more challenging intended purposes, and
- More able persons are more likely to receive higher scores on all ESI items and intended purposes than are less able persons.

Again, when we analyzed the data for the persons in the ESI standardization sample, we found that 87.2% of the persons demonstrated goodness of fit to the MFR model of the ESI (infit and outfit $MnSq \leq 1.4$, $z < 2.0$). Among the 146 persons who failed to demonstrate acceptable goodness of fit, 40 (27%) were associated with one rater. When that rater's data were omitted from consideration, the ***overall level of goodness of fit for persons increased to 90.4%, supporting person response validity of the ESI.*** While up to 10% person misfit is common among performance-based assessments, we anticipate that the rate of acceptable person fit will increase following the revisions we have made with regard to the items *Discloses* and *Thanks*.

3.9 Evidence of Ability to Measure Influences of the Primary Social Partner on Quality of Social Interaction

In an earlier study of the ESI, Asplund and Forsberg (2006) included familiarity of the social partner as an additional facet in their MFR analysis of their ESI data (see Section 2.2.2, Chapter 2). By doing so, they were able to evaluate if there was a significant uniform impact of partner familiarity on a person's ESI quality of social interaction measure. Their finding that interaction with an unfamiliar social partner was easier (0.32 logit difference) than was interaction with a familiar social partner can also be viewed as evidence that the ESI quality of social interaction measures are *sensitive enough to evaluate the impact of partner familiarity on a person's quality of social interaction*.

Including partner familiarity as a facet, however, also has a potential negative consequence — rather than developing a measure that can be used to evaluate differences in a person's quality of social interaction with different social partners, the external influences become obscured. Therefore, we have deliberately chosen not to account for such influences in the estimation of a person's ESI quality of social interaction measure. Instead, *we propose using the ESI in research designed to study and understand what features of the physical and social environment most support or limit a person's quality of social interaction*.

To generate preliminary evidence (sample size = 468 persons) that the ESI could be used successfully in such research, we used MFR analysis to evaluate what characteristics of the primary social partner most influenced a person's quality of social interaction. When familiarity of the primary social partner, status of the primary social partner, and primary social partner's overall quality of social interaction were considered simultaneously, *partner familiarity* had minimal impact — quality of social interaction is likely to be only slightly better when the primary partner is someone who is *familiar* or *somewhat familiar* than when the social partner is someone who is *unfamiliar* (maximum difference = 0.10 logit).

When we analyzed data for the current standardization sample ($n = 1140$ persons), we obtained similar results. That is, the impact of familiarity of the social partner was minimal, and social interaction with someone who was somewhat familiar

was slightly better than was social interaction with someone who was familiar or unfamiliar (maximum difference = 0.12 logit)

In the earlier analysis, the feature of the primary social partner that appeared to have an even greater impact on a person's quality of social interaction is the **primary partner's status** (maximum difference = 0.24 logit). This was confirmed in the analysis of the data for the current standardization sample (maximum difference = 0.23 logit). That is, a person's ESI quality of social interaction measure is most likely to be lower if the primary social partner is a *family member/relative*. The social partner statuses that are least likely to impact a person's ESI quality of social interaction measures are those of *expert/supervisor/teacher/service provider, friend/colleague/classmate, and other acquaintance*. Finally, the social partner statuses associated with higher ESI quality of social interaction measures are those of *receiver of services/customer*.

It is the **primary social partner's overall quality of social interaction**, however, that appears to have the greatest impact on a person's ESI quality of social interaction measure (maximum difference = 0.27 logit). As would be expected, the person's ESI measure is more likely to be progressively higher as the primary social partner's overall quality of social interaction progresses from being markedly inappropriate to appropriate.

When considered simultaneously, the primary social partner characteristics having the least to most impact on a person's quality of social were:

- **Familiarity of the primary social partner**
- **Status of the primary social partner**
- **Primary social partner's overall quality of social interaction**

3.10 ESI Item Difficulty and Intended Purpose Challenge Hierarchies

The ESI item difficulty and intended purpose challenge calibration values based on the data for the 1140 persons included in the ESI standardization sample are shown in Table C–1, Appendix C. The difficulties of the ESI items ranged from 1.16 logits to -1.85 logits, and the intended purpose challenges ranged from 0.21 to -0.10 logit.

These item difficulty and intended purpose challenge calibration values have been incorporated into the ESI computer-scoring program. Once a potential ESI rater has calibrated successfully and has obtained a rater calibration code that can be entered into the software program, the calibrated ESI rater can generate ESI reports based on many-faceted Rasch analyses of the data for the persons he/she has tested using the ESI (see Section 8.1, Chapter 8 and Section 10.2 and 10–3, Chapter 10 for more information about computer-generated ESI reports).

Rater calibration refers to the process of determining an occupational therapist's degree of leniency or severity when rating the ESI items during an ESI training course, as well as determining the degree to which the occupational therapist scored the ESI in a valid and reliable manner when testing persons after completion of the training course.