THE TERM RELIABILITY

The primary goal of this training series is to teach clinicians to produce reliable clinical data using two-way scoring, five-way scoring, and phonetic transcription. What are reliable data?

Reliable data, like reliable friends, can be depended on. Specifically, reliable data can be depended on to agree with some criterion measure, such as other data taken by the same person or data taken by a criterion judge. The degree of reliability or dependability, however, is entirely a matter of practicality. Just as a person’s being a few minutes late may be tolerable in certain situations, a mere “ballpark” estimate of a particular phenomenon may be sufficient for making certain clinical decisions. In other situations, however, a person’s dependability within a few seconds may make the difference between life and death. The point is that the degree of reliability or dependability—the amount of agreement we require between our data and criterion measurements—is entirely an individual matter. To illustrate this point, consider the following clinical situation.

Suppose we have a child learning to say /s/ by repeating a list of ten words containing an /s/ target. Every time the child says /s/ correctly, as judged by the clinician, the child gets a point toward some favorite play activity. In this clinical situation, how reliable does the clinician need to be in making two-way (“correct–incorrect”) scoring decisions? 100 percent? 95 percent? 90 percent? 85 percent? How critical is it that the clinician reinforce all and only correct /s/’s in order for the child to acquire the correct /s/?

This question has never been studied. However, most learning theorists would agree that less than 70 percent reliability would not be fair to the child in this learning situation. At the upper end, however, 100 percent reliability may not be necessary either. Notice that if only ten items were being trained, 70 percent reliability would mean that to three of the ten responses the child makes, the clinician would say “correct” when the child was really “incorrect,” or vice versa. Would a child acquire a correct /s/ under these feedback conditions? Again, it depends on the situation. The degree of reliability required of a person or of an instrument depends entirely on permissible tolerance under particular conditions.

Measurements made by means of a micrometer are more precise than those made with the common ruler; however, each tool has its own error factor in terms of its quality of construction and in terms of the person doing the reading. Similarly, although five-way scoring may be more precise than two-way scoring, each system has its own error factors. The next section will show you some ways to assess your own error factor so that you can calculate your dependability, when using the three systems presented in this training series. The last section summarizes research findings on factors that influence the reliability of clinical transcription.

PROCEDURES TO CALCULATE TRANSCRIPTION RELIABILITY

Four Types of Reliability Checks

Table D-1 is a representation of the four types of reliability checks that should be familiar to clinicians.

Intrajudge agreement, how well a person agrees with himself or herself, can be assessed in two ways. As indicated in Table D-1, one way is to record a live session; scoring or transcription done later from the recording is compared to that which was done live. This way of assessing reliability is efficient for the working clinician because it requires the least amount of time. As described in Chapter 7, however, comparison of judgments made live with those made from recordings introduces uncontrolled factors.

The second type of intrajudge assessment (Table D-1), recording-to-recording, compares responses made to the same recording scored on two separate occasions. This controls the stimulus conditions, as long as comparable playback conditions are maintained.

Interjudge agreement, agreement between two or more persons, can be assessed either live or from a recording, as described in Table D-1. In this series, your interjudge agreement is assessed each time you compare your scores of transcription to the key. Your agreement with the consensus key is a measure of how often you agree with the “experts.” If you assume that our keys are valid (!), your relative compe-


How to Calculate Percentage of Agreement

Reliability of transcription is assessed as the percentage of agreement with an established source. As just discussed, this other source can be the same examiner, another examiner, or a group of examiners. Obtaining a percentage of agreement is not a difficult task. The formula is:

\[
\text{Percentage of agreement} = \frac{\text{Number of units scored similarly}}{\text{Total number of units scored}} \times 100
\]

In this general formula, the “units” used are defined in terms of the variable of interest. For example, units could be all /s/ sounds, all fricatives, all /t/ sounds, or all dentalized /s/’s. Whichever unit is chosen, it is the comparison between each established criterion unit and each comparison judgment that is of interest.

How to Calculate Reliability in This Text

We now confront the issue of how much reliability we should expect when people score or transcribe disordered speech. We know that 100 percent agreement, whether intrajudge or interjudge, is not realistic to expect, nor is it generally necessary to achieve. In this section, we will provide some guidelines for how to compare your scoring and transcription of the recorded materials to the answer keys provided in this manual. Two concepts underlie the strategies we suggest.

The first concept is that of functional equivalence between different phonetic transcriptions. In this situation, two transcribers may perceive similar speech events but elect to use somewhat different phonetic symbols to capture this essentially similar perceptual event. If one transcriber symbolizes a sound as [i] (lowered /i/) and the other transcriber hears it as [ɪ] (raised /ɪ/), do they really disagree? For all but the most fine-grained analyses, the two transcriptions are functionally equivalent. Now consider a slightly different situation, the transcription [s] versus [tʃ]. These transcriptions are nearly functionally equivalent, in that each symbolizes a frictionalized sound made in the area of the dental-alveolar ridge. Certainly, these two symbols are more nearly equivalent than, for example [s] versus [ʃ]. These examples illustrate the point that transcriptions may “agree” to different degrees: (1) They may be identical, (2) they may be functionally equivalent, or (3) they may be nearly functionally equivalent.

The second and related concept concerns the arbitrary nature of reliability criteria, discussed earlier in terms of a person’s dependability. In some measurement situations, we can justly afford to set a liberal (low) standard of agreement. In other situations, we must set a more strict (high) criterion for an acceptable level of agreement. Specifically, in some situations, transcriptions that are functionally equivalent or nearly functionally equivalent may be usefully considered to be in agreement. In other situations, only identical transcriptions should be scored as agreements. Consider three transcriptions of a child’s response for the word pan:

Transcriber 1  [p əʊ n ]
Transcriber 2  [p æ ]
Transcriber 3  [p ə ]

Does Transcriber 2 agree with Transcriber 1 on the final segment? We would have to say by a strict criterion that they do not agree. Transcriber 1 has used a final /n/, whereas Transcriber 2 has not. Transcriber 3 also disagrees with Transcriber 1 in that /n/ is not present, but Transcriber 3 has symbolized the vowel as nasalized. By a strict criterion, each transcriber disagrees with the others on the final segment. But by a more liberal criterion, Transcriber 3, who has marked nasalization on the vowel, may be considered to be more in agreement with Transcriber 1 than Transcriber 2 is.

---

### Table D-1: Four Types of Reliability Assessment

<table>
<thead>
<tr>
<th>Type of Reliability</th>
<th>Assessment Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrajudge</td>
<td>Live—Recording&lt;br&gt;One clinician scores/transcribes both live and from a recording of the same session and compares the results.</td>
</tr>
<tr>
<td>Interjudge</td>
<td>Live—Recording&lt;br&gt;Two or more clinicians score/transcribe the same session live and compare their results.</td>
</tr>
<tr>
<td></td>
<td>Recording—Recording&lt;br&gt;One clinician scores/transcribes the same recording on two occasions and compares the results.</td>
</tr>
<tr>
<td></td>
<td>Recording&lt;br&gt;Two or more clinicians score/transcribe a recording at same or different times and compare their results.</td>
</tr>
</tbody>
</table>
because Transcriber 2 has marked neither the nasal segment nor the nasalization on the vowel. Because we know that hearing word-final /n/ can be difficult, we may elect to call the transcriptions of Transcriber 1 and Transcriber 3 nearly functionally equivalent on the basis of their vowel transcriptions.

You can use these two concepts—functional equivalence and setting a strictness criterion—in calculating your agreement with the answer keys. Each time your transcription differs from that given in the key, ask yourself whether your transcription is functionally equivalent or nearly functionally equivalent with that provided in the key. You may then want to calculate your percentage of agreement both in terms of a strict criterion (perhaps identical transcription with that given in the key or in the consensus column) and also by a more liberal criterion (functional or near functional equivalence). To parallel the previous example, the fact that you recognize the presence of a stop—although you may disagree with the key on the exact nature of the stop—should be more highly valued than missing the stop altogether. These matters should be discussed with your instructor.

To summarize, for any given speech event, there are several ways to transcribe its major characteristics. Particularly in clinical transcription, which uses many diacritic markings, alternate transcriptions of an event may be functionally equivalent. You can be “right on,” “close,” “in the ballpark,” or “off by a mile.” To get useful feedback on your learning, you should calculate your percentage of agreement with the experts using both strict and liberal criteria. In clinical situations, you will need to keep aware of how dependable you are in your phonetic skills. Just as you need to calibrate periodically an instrument such as an audiometer—to ensure validity and reliability of measurement—you also need to check regularly the reliability of your scoring and phonetic transcription.

RELIABILITY FINDINGS IN CLINICAL TRANSCRIPTION

Chapter 7 introduced several factors associated with the reliability of phonetic transcription in clinical settings. Because transcription is basic to the validity of assessment and management, the reliability of phonetic transcription has been the source of many studies. For your interest as a student of phonetics, we provide two summaries of this research.

The first research review, Table D-2, summarizes findings sampled from over a dozen studies crossing a variety of clinical settings. The right-most column of conclusions provides wide-ranging perspectives on the reliability of phonetic transcription. You may find it interesting to review some of these generalizations with your instructor, perhaps arranging to pursue your own research for an extra credit project. References for all papers cited in Table D-2, plus some core references in phonetic transcription, are included in the bibliography at the end of this appendix.

A second research review, Table D-3, is taken from a series of studies conducted in one research setting (Shriberg and Lof, 1991). These reliability findings were obtained from five two-person phonetic transcription teams. Each team learned to do phonetic transcription using the symbols and procedures described in Clinical Phonetics. They also learned procedures to reach a consensus between their individual judgments (Shriberg et al., 1984). As you look at the generalizations from these studies, you will see that not all teams were able to develop clinically reliable narrow phonetic transcription. Specifically, the interjudge agreements of some teams (i.e., their consensus agreements assessed on the same recorded speech samples at two different times) was less than 70 percent.

The challenge of increasing intrajudge and interjudge agreement has been the source of several studies and reviews in the years since the data summarized in Tables D-2 and D-3. Essentially, estimates of transcription agreement on consonants, vowels, and especially diacritics continue to be often unacceptably low, especially among researchers and speech-language pathologists with more limited training and with children who have significant speech involvements (e.g., Cucchiarini, 1996; Grunwell and Harding, 1996; Ingrisano, Klee, and Binger, 1996; Maassen, Offerenga, Viergge, and Thoonen, 1996; Vierregge and Maassen, 1999; Gooch, Hardin-Jones, Chapman, Trost-Cardamone, and Sussman, 2001; Howard and Heselwood, 2002). Moran and Fitch (2001) have recently demonstrated that phonetics students’ phoneme awareness, as assessed by a battery of measures, may be significantly associated with their scores on transcription quizzes. These authors speculate on the possibility of improving transcription skills by providing students with activities designed to increase their phoneme awareness. Such strategies, as well as the availability of new transcriptions systems and audio training programs, may help in teaching skills that may increase the reliability of clinical transcription (e.g., Ball, Code, Rahilly, and Hazlett, 1994; Ball, Rahilly, and Tench, 1996; Small, 1999; Ball, Esling, and Dickson, 2000; Hoffman and Buckingham, 2000; Louko and Edwards, 2001; Pollock and Berni, 2001; Powell, 2001; Snow, 2001; Stoel-Gammon, 2001). Moreover, it is likely that substantial improvement in transcription agreement may be achieved using procedures that provide online acoustics information (see Chapter 10) on segmental and suprasegmental aspects of speech production. Until such potentially effective systems become readily available, however, it is important to acknowledge the reliability challenges associated with broad and narrow phonetic transcription as an auditory-perceptual skill.
<table>
<thead>
<tr>
<th>Source</th>
<th>Variable</th>
<th>Study</th>
<th>Mean Agreement Findingsa</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Subjects</td>
<td>1. Intelligibility</td>
<td>Philips and Bzoch (1969)</td>
<td>Interjudge variability with intelligibility; ( r = 0.19 )</td>
<td>There is little association between interjudge agreement and intelligibility.</td>
</tr>
<tr>
<td></td>
<td>2. Severity of involvement</td>
<td>Irwin (1970)</td>
<td>Interjudge: correct sounds = 88%; misarticulated sounds = 66%</td>
<td>Agreement is considerably higher when calculated on correct, compared to misarticulated, sounds.</td>
</tr>
<tr>
<td></td>
<td>3. Type of error</td>
<td>Philips and Bzoch (1969)</td>
<td>Agreement on error type (five classifications): intrajudge = 50–91%; interjudge = 6–19%</td>
<td>Judges have low levels of agreement on each of five categories of error classifications.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Norris, Harden, and Bell (1980)</td>
<td>—</td>
<td>Interjudge agreement on omissions was higher than on substitutions.</td>
</tr>
<tr>
<td></td>
<td>4. Clinical significance</td>
<td>Shriberg, Kwiatkowski, and Hoffmann (1984)</td>
<td>Exact retest consensus reliability was 68%; with nonerror diacritics removed, exact agreement was 76%.</td>
<td>Consensus transcription agreement is higher when based only on sound changes that have clinical significance.</td>
</tr>
<tr>
<td>B. Analyses</td>
<td>5. Transcribers</td>
<td>Siegel (1962)</td>
<td>Interjudge: ( r = 0.92 )</td>
<td>Transcribers can be trained to high agreement levels in making correct/incorrect judgments of sounds in isolated words, but differences on scores assigned by individual transcribers can be considerable.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Burkowsky (1967, 1971)</td>
<td>Intrajudge = 64%; interjudge = 36%</td>
<td>“The field of speech pathology does not produce students with proven competence in listening to speech sound production” (p. 1).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Irwin (1970)</td>
<td>Interjudge = 87%</td>
<td>“Undergraduate majors in speech pathology were relatively [reliable]” (p. 554).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diedrich and Bangert (1976, 1981)</td>
<td>—</td>
<td>Sounds may be judged as correct more often by judges who also are functioning as the child’s clinician.</td>
</tr>
<tr>
<td></td>
<td>6. Type of agreement</td>
<td>Schissel and Flournoy (1978)</td>
<td>—</td>
<td>Intrajudge agreement was higher than interjudge agreement for both experienced and inexperienced listeners.</td>
</tr>
<tr>
<td></td>
<td>7. Type of system</td>
<td>Amorosa, von Benda, Wagner, and Keck (1985)</td>
<td>Interjudge: live = 56%; tape = 72%</td>
<td>Transcription procedures that do not allow for unlimited replays will result in over-diagnosis of phonologic disability because “all other information on phonetic detail has either been omitted or must be considered unreliable” (p. 286).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pye, Wilcox, and Siren (1988)</td>
<td>—</td>
<td>Broad transcriptions were used because “the frequency with which two or more individual transcribers chose to use the same diacritic marker for the same segment was quite small” (p. 21).</td>
</tr>
<tr>
<td>C. Contexts</td>
<td>8. Agreement criteria</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>9. Sampling mode</td>
<td>Irwin and Krafchick (1965)</td>
<td>—</td>
<td>Identification of misarticulations was better in words than in phrases.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pye, Wilcox, and Siren (1988)</td>
<td>—</td>
<td>Interjudge agreement was higher for articulation test responses than for continuous speech.</td>
</tr>
<tr>
<td></td>
<td>10. Structural, grammatical, and stress forms</td>
<td>McCauley and Skenes (1987)</td>
<td>Significantly more unstressed than stressed /r/’s scored as correct</td>
<td>(Among other interpretations), listeners may have a more lenient standard for correct /r/ in unstressed, compared to stressed, contexts.</td>
</tr>
<tr>
<td></td>
<td>11. Word position</td>
<td>Philips and Bzoch (1969)</td>
<td>Interjudge: word-initial = 80%; word-medial = 78%; word-final = 67%</td>
<td>“. . . sounds in final positions account for a greater portion of the disagreements” (p. 28).</td>
</tr>
<tr>
<td></td>
<td>12. Target environment</td>
<td>Ruscello, Lass, Posch, and Jones (1980)</td>
<td>6–24% judgment shifts on correct, moderate errors, and severe errors on /r/ and /s/</td>
<td>“. . . alterations in judgment . . . occur when individuals listen repeatedly to the same stimuli” (p. 5).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Norris, Harden, and Bell (1980)</td>
<td>—</td>
<td>Syllabic function contributed little to interjudge agreement.</td>
</tr>
</tbody>
</table>

*aBlank entries in this column reflect situations in which single agreement figures would not be representative.*
<table>
<thead>
<tr>
<th>Source</th>
<th>Variable</th>
<th>Sample Levels</th>
<th>Generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Subjects</strong></td>
<td>1. Intelligibility</td>
<td>High, medium, low</td>
<td>Transcriber agreement on consonants and vowels has a low to moderately positive association with the subjects’ severity of involvement, as indexed by percentage of consonants correct and intelligibility.</td>
</tr>
<tr>
<td></td>
<td>2. Severity of involvement</td>
<td>Mild, moderate, severe</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Type of error</td>
<td>Deletion, substitution, distortion</td>
<td>Neither the absolute nor relative percentages of each of the primary error types—deletions, substitutions, or distortions—are highly associated with transcription agreement.</td>
</tr>
<tr>
<td></td>
<td>4. Clinical significance</td>
<td>Articulation error, acceptable allophone</td>
<td>Transcriber agreement is not associated with the clinical significance of diacritical description of speech.</td>
</tr>
<tr>
<td><strong>B. Analyses</strong></td>
<td>5. Transcribers</td>
<td>Background, training</td>
<td>No generalization</td>
</tr>
<tr>
<td></td>
<td>6. Type of agreement</td>
<td>Intrajudge, interjudge, consensus</td>
<td>The two types of transcription agreement—interjudge and intrajudge—have essentially similar average percentages of agreement, ranging from the mid-60s to the mid-high-90s.</td>
</tr>
<tr>
<td></td>
<td>7. Type of system</td>
<td>Broad, narrow (International Phonetic Alphabet, other)</td>
<td>The two systems of phonetic transcription—broad (93%) and narrow (74%)—differ in average transcription agreement by approximately 20 points.</td>
</tr>
<tr>
<td></td>
<td>8. Agreement criteria</td>
<td>Exact, within-class, other</td>
<td>The three types of transcription agreement criteria for diacritics—exact (33%), within-class (40%), and any diacritic (48%)—differ in average transcription agreement (uncorrected for chance agreement) by a range of approximately 15 points.</td>
</tr>
<tr>
<td><strong>C. Contexts</strong></td>
<td>9. Sampling mode</td>
<td>Continuous speech, articulation test</td>
<td>Transcription agreement based on continuous speech samples is somewhat higher than agreement based on articulation test responses.</td>
</tr>
<tr>
<td></td>
<td>10. Structural, grammatical, and stress forms</td>
<td>Canonical, grammatic, stress</td>
<td>No generalization</td>
</tr>
<tr>
<td></td>
<td>11. Word position</td>
<td>Initial, medial, final</td>
<td>Of the three word positions, word-initial consonants are generally transcribed most reliably, with word-final consonants typically associated with the lowest reliability.</td>
</tr>
<tr>
<td></td>
<td>12. Target environment</td>
<td>Stimulus context, phonetic context</td>
<td>No generalization</td>
</tr>
</tbody>
</table>

*Shriberg and Lof, 1991.*
### TABLE D-3 (Continued)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variable</th>
<th>Sample Levels</th>
<th>Generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Units</td>
<td>13. Class</td>
<td>Consonants, vowels</td>
<td>Average transcription agreement at the level of phonetic features and classes is within acceptable levels for broad transcription and generally below acceptable levels for narrow phonetic transcription.</td>
</tr>
<tr>
<td></td>
<td>14. Features</td>
<td>Manner, place, voicing, height</td>
<td>The reliability of broad transcription of vowels in a sample is essentially independent of their rank order of occurrence and percentage correct. For consonants, transcription agreement is independent of rank order of occurrence and lower, but within an acceptable range, for the 12 most frequently misarticulated sounds.</td>
</tr>
<tr>
<td></td>
<td>15. Sounds</td>
<td>24 consonants, 17 vowels/diphthongs</td>
<td>Average transcription agreement percentages for each of the 41 sounds are within acceptable levels for broad transcription but generally below acceptable ranges for narrow phonetic transcription.</td>
</tr>
<tr>
<td></td>
<td>16. Diacritics</td>
<td>35 symbols for narrow transcription</td>
<td>There are substantial differences in the average number of diacritics per word used by different consensus transcription teams within and between sampling modes and subject groups. There is fairly stable consistency in the average number of diacritics per word used by the same consensus transcription team doing narrow phonetic transcription on the same speech sample. The proportional occurrence of individual diacritic symbols in narrow phonetic transcription ranges from low to moderately high depending on consensus transcription teams, subject groups, and sampling modes. Transcription agreement on an individual diacritic is essentially independent of its proportional occurrence in a speech sample. The average interjudge and intrajudge percentage of agreement estimates for diacritic transcription are below acceptable reliability boundary levels, even at the least strict agreement criteria.</td>
</tr>
</tbody>
</table>
Glossary

**Functional equivalence** essentially equivalent phonetic transcriptions of a target behavior that uses alternative symbolization.

**Interjudge agreement** the extent to which two or more clinicians make similar judgments about a target behavior.

**Intrajudge agreement** the extent to which a clinician makes similar judgments on different occasions about the same target behavior.

**Near functional equivalence** nearly equivalent phonetic transcriptions of a target behavior in terms of place and manner features.

**Percentage of agreement** the number of target units scored or transcribed similarly divided by the total number of units scored (multiplied by 100).

**Strictness criteria** an arbitrary basis for determining agreement between transcriptions. **Liberal criteria** might allow transcriptions that are nearly functionally equivalent to be counted as agreements; **strict criteria** might allow only identical transcriptions to be counted as agreements.

Exercise

1. To assess her intrajudge reliability, a clinician gives a child an articulation test on two separate occasions. She then compares her score sheets taken from the two sessions. Is this a good way for the clinician to assess her intrajudge reliability? Why or why not?

2. To assess his interjudge reliability, a clinician scores a child live, then brings a recording of the session to his clinical supervisor. The clinician then calculates a percentage of agreement between his scores and those of his supervisor for all /s/ words in the sample. Is this a correct way for the clinician to assess his interjudge reliability? Why or why not?

3. A clinician scores a recorded audio sample on two separate occasions and finds that she agrees with herself on 22 out of a possible 25 items. What is her percentage of intrajudge agreement?

4. A clinician scores a recording of a child on two separate occasions. He finds that he agrees with himself on 22 items and disagrees on 8 items. What is his percentage of intrajudge agreement?

5. On a 67-item test scored by two examiners, Examiner 2 disagrees with Examiner 1 on 11 items. What is the percentage of agreement between Examiner 1 and Examiner 2?

A bibliography on transcription reliability


**ADDITIONAL TRANSCRIPTION RELIABILITY REFERENCES FOR CLINICAL PHONETICS, 4TH EDITION**


Johnson, J. M., Munson, B., & Edward. J. (June, 2010). The role of clinical experience in listening for phonetic detail in children’s speech. Poster presentation at the Symposium for Research in Child Language Disorders, Madison, WI.


