Issues in psychometry and child testing

Clay Williams
Akita International University

Abstract
While psychometric testing techniques often can provide the most precise and verifiable data regarding internal language processing, many of the standard reaction time testing paradigms face serious issues in their use with young learners (i.e., elementary school ages and below). These issues include such problems as slower reaction times, decreased attention span, slower reading ability, difficulty in understanding the task, etc. While there have been successful adaptations of certain psychometric techniques (e.g., length of gaze studies and eye-tracking techniques), many practitioners of psychometric research are reluctant to test children using standard testing paradigms such as priming tests and reaction time studies, despite the wealth of information on language acquisition which child testing could potentially reveal. This paper will draw from first-hand experience of studies on young learners and reveal the problems which accompany the application of psychometric testing methods to children, as well as how to overcome these issues. Viable techniques for child testing according to age groups will be discussed, along with practical considerations and advice for successful use of reaction time tests and other psychometric testing paradigms with young subjects.

1. Introduction: Psychometric testing

In the veritable panoply of available research methodologies available to applied linguists, psychometry is often singled out as the branch of inquiry most representative of the classical experimental method and of quantitative approaches to inquiry. Chaudron (1988; as cited in Nunan, 1992) argues that applied linguistics research could be subdivided into the following four research types: 1) psychometric; 2) interaction analysis; 3) discourse analysis; and 4) ethnography. While such a simplistic division of research methodologies may quite rightly invite criticism as being overly reductive, nevertheless, it does demonstrate that psychometric methodologies occupy an enormous space in the applied linguistics literature. Psychometry consists of a broad range of testing methodologies which were borrowed from psychological study techniques and are invariably quantitative in nature.

1.1 Reaction timing

At the heart of psychometry is the reaction time experiment. In this type of experiment methodology, the subject reaction to specific stimuli (usually by pushing a button) is timed (usually to the millisecond). The explicit assumption of the methodology is that the reaction time allows the researcher to gauge mental processing time, and that significant differences in reaction time between different stimuli are caused by processing differences. Welford (1980) categorized reaction time experiments into four categories: 1) Simple reaction timing (e.g., there is a single stimulus and a single response type): examples of this would include having subjects push a button when they see a dot on the screen. The latency time between the appearance of the dot and the subject pushing the button would be measured. 2) Recognition reaction timing: here there are two different types of stimuli – those which should be responded to, and those which should not be responded to – and therefore the subject is confronted with a dual, yes/no-style choice. A common example of this type of test would be the lexical decision test which is a mainstay test procedure among applied linguists. In this type of research,

subjects are presented with letter strings, and the subject must decide whether or not the letter string constitutes a word by hitting one of two buttons signifying a "yes" and a "no" response, respectively. 3) Choice reaction timing: in this methodology there is a wide array of possible responses, but the response must match the stimulus. For example, the subject could be tasked with typing the letters that appear randomly on the screen. 4) Serial reaction timing: this experiment type is much like "choice" in that there is a wide array of response types; however, unlike in the "choice" dynamic the stimuli are not randomized, which leads to subject improvement and faster response over time (e.g., if the letters subjects were typing were spelling words, word recognition would enable the subject to get faster as one approached the end of the word, as the subject would be intuiting what letters come next). Mean reaction times have been shown to depend greatly on the task type. At its most basic, simple reaction time experiments tend to elicit the quickest response, followed by recognition reaction time experiments, which are followed in turn by choice reaction times (e.g., Laming, 1968; O'Shea & Bashore, 2012). Obviously, the relative complexity of the task correlates directly with the amount of processing time. Stimulus complexity, as well, has been shown to slow reaction speed (Luce, 1986). For example, word length directly impacts recognition speed (i.e., long words are recognized more slowly than short words).

While there is a wide variety of methodological variations in stimulus presentation, in linguistics research, recognition testing procedures seem to be the most- commonly employed. In addition, testing can make use of sub-task-level variables such as priming techniques. Priming is the use of one stimulus to either facilitate or inhibit the activation of another stimulus (e.g., if one first sees the word "nurse," one would subsequently recognize the word doctor measurably more quickly). It comes in a variety of types, such as repetition priming (i.e., repeating the same word twice), interval priming (i.e., where the word is repeated, but not immediately – i.e., different word targets occur before the repetition), and masked priming (i.e., using a forward mask and a presentation time of <80ms, the prime is rendered invisible to the subject's conscious mind; however, the subject will still react to the presence of the prime).

While psychometric study techniques have been a mainstay of applied linguistic research for many decades, there are limits on their applicability to certain subject populations. This paper will investigate the potential problems one will encounter in using traditional psychometric reaction timing experimental techniques on child subjects.

2. Reaction timing with child subjects

It must be widely acknowledged that, due to both physical and cognitive maturational issues, children, in many ways, constitute a very different subject group as compared to adults. This holds true in practically any type of experimental endeavor, and it certainly holds true for the field of applied linguistics. On the one hand, children's unique language abilities are rightly the subject of focus for linguistic inquiry. The rapid period of L1 acquisition and development in early childhood, as well as the way in which young children go about learning L2s (in many ways more reminiscent of L1-learning than of how adults typically learn L2s), are vital areas of linguistic research, and any claim to knowledge of how human language ability works that ignores these unique childhood factors would likely fall far short of grasping the full span of the marvel which is human language. This principle of the uniqueness of children holds true as well for the relative usability of different applied linguistics experimental methodologies. Reaction timing, and psychometric data in general, is often a poor fit for collecting reliable data with child subjects. In short, certain inherent properties common to children can limit or completely nullify the value of data yielded through psychometric testing means. The main issues include slower overall reaction times, slower reading speeds, and lower attention spans.

2.1 Child reaction speed

Due to both physical and psychological maturational factors, there is a significant decrease in reaction timing throughout childhood. Bucsuházy and Samela (2017) found that simple reaction times (i.e., pushing a button in response to a stimulus) for children varied significantly throughout childhood: ages 3-5 > ages 6-7 = ages 8-9 > ages 10-14 > ages 15-18 = ages 20-30. While the overall reaction time of 6–7-year-olds did not differ significantly from 8–9-year-olds, the variation in reaction time was broader for the younger children. As tests were made more complex, a steady, inverse, significant relationship between age and reaction times held up until age 15. This degree of reaction speed difference can by itself effectively eliminate any productive intergroup comparisons across ages (i.e., subjects must be grouped very closely in age to be able to compare RTs).

2.2 Child reading speed

This issue does not constitute a great mystery. As children are typically still actively engaged in the task of mastering L1 literacy skills, their overall reading speeds tend to be significantly lower than that of adults, which will invariably affect the overall reaction timing of any reading-based tests such as lexical decision or word recognition tasks. Furthermore, this variability in reading speed is potentially greater even than that of reaction speed, and children do not "catch up" to typical adult rates of reading until after high school (which reflects the impact of continuing tertiary education on adult reading speeds). Reading speed per grade level are shown below in Table 1.

Table 1. Reading speeds across grade levels

1st Grade: 53-111 words per minute (WPM)	5th Grade: 139-194 WPM
2nd Grade: 89-149 WPM	6th-8th Grade: 150-204 WPM
3rd Grade: 107-162 WPM	High School: 200-300 WPM
4th Grade: 123-180 WPM	Adult: 220-350 WPM

(Adapted from Hasbrouck & Tyndal, 2017)

2.3 Child attention span effects

Getting subjects to pay close attention to test stimuli can be challenging with subjects of any age. When the researcher is attempting to extract millisecond-level reaction time data, anything – such as a sudden noise in the next room or an insect flying around the testing facility – can potentially distract subjects long enough to nullify the results. The need for maintaining subject attentiveness is why many researchers make use of performance incentives in their research (e.g., giving subjects some additional renumeration if they score at a predetermined high level of accuracy). While getting even adult subjects to maintain concentration during testing can be difficult enough, the limited attention span of young children can directly complicate standard testing protocols. For example, the usual recommendation for collecting RT data is to conduct a full practice period (for the subject to get accustomed to the task type and the user interface), and then to collect upwards of 300 reaction times per participant (e.g., Sanders, 1998). Depending upon the task-type, this could easily mean anywhere from 60-90 minutes of testing time, and one can immediately intuit how such a long testing period could result in small children losing focus.

3. A case study example of child psychometric testing

To exemplify the sorts of issues faced by researchers when attempting to conduct investigations using young children as subjects, the author will present a case study based upon research conducted from 2016-2020 (e.g., Williams & Naganuma, 2018; 2020). This study, whose central aim was to investigate how child L2 vocabulary acquisition differed from that of adults, and specifically whether children have the unique ability to create instantaneous conceptual links to L2 words, serves as an object lesson in types of issues often faced when testing young learners.

3.1 Subjects

The initial testing to calibrate test instruments and procedures (which will be our focus for this case study) was conducted on 63 students in two Japanese elementary schools with ages ranging from the 2nd grade (i.e., 7-8 y.o.) to 6th grade (i.e., 11-12 y.o.). The breakdown by grade was six 2nd graders, nine 3rd graders, eighteen 4th graders, eleven 5th graders, and nineteen 6th graders. The considerable variance between group sizes was reflective of class size differences. All subjects spoke Japanese as their native language, and none of the tested children spoke any other language at home. The 5th and 6th grade children received one 45-minute period per week of oral English instruction, pursuant to Japanese law. The other children received periodic, unstructured English activities (i.e., games and such), but no formal instruction.

3.2 Materials

The test was delivered via the DMDX platform (Forster & Forster, 2003). In the test, the subject would hear an English word presented auditorily (via headphones). The sound stimulus was immediately followed by the presentation of either two pictures or two words written in Japanese. The subjects were to match the English target word with the corresponding picture or Japanese translation by hitting one of two marked buttons, corresponding to the right-hand and left-hand sides of the screen. There were a total of 20 auditory word prompts preceded by 6 practice items.

3.3 Analysis

Reaction times were automatically recorded by the DMDX software, and later those times were analyzed by age groupings.

3.3 Results

The collected data, once analyzed, revealed that 2nd and 3rd graders' reaction times averaged much longer than that of the older children. Second graders averaged 1961ms to respond, and 3rd graders averaged 1737ms. By comparison, 4th graders averaged 1362ms, 5th graders averaged 1125ms, and 6th graders averaged a comparable 1117ms.

3.4 Discussion of results

What immediately leaps to the forefront of the reader's attention about these results is the noticeable "jump" in word recognition speed starting from 4th grade, and leveling off in 5th grade. As we've seen previously from the Bucsuházy and Samela (2017) study on reaction timing by age, physical reaction time maturational differences are not sufficiently distinct at these ages (roughly 7-9 years old) to cause such profound distinction in the reaction times for this study.

Fortunately, the culprit was easily identified as, during the actual data collection, there were quite a few cases of 2nd graders either reading the Japanese translation stimuli aloud before responding, or even asking testers what certain words were. Here, the cause of this slowdown was the differential in reading speed. Because the 2nd graders (and 3rd graders, to an admittedly lesser degree) were still developing their basic word recognition skills, the RTs being captured by the test were not reflective of the time it took for them to recognize the English word and to connect it to the Japanese translation. Instead, the reaction time became, in large part, a measure of their reading speed. In this case, the easy solution to the problem which would enable the research to continue unencumbered was to eliminate 2nd and 3rd graders from further analysis, and as the calibrations ended and actual testing commenced, the study would thereafter focus solely on 4th-6th graders.

4. How to study children using psychometric techniques

Obviously, the presence of such significant variations in RT both between age groups in children, as well as between children and adults, can make psychometric testing methods ill-suited for studying certain language acquisition phenomena in children. Certainly, the best overall advice for testing children would be to simply avoid testing methods whose results would be directly affected by age variables. Fortunately, there are certain psychometric techniques which can still be used with children without fear of significant differences stemming from motor skills or psychological maturational issues contaminating results. For instance, eye-tracking studies and length of gaze measures are relatively unaffected by such variables and can be particularly effective in studying very young (i.e., pre-verbal) children. Still, eye-tracking hardware can be expensive, which can preclude its use among researchers not affiliated with well-funded and well-equipped laboratories. Fortunately, by following a few basic principles, one can still make productive use of some (low-cost) psychometric study methodologies.

4.1 Avoid RT for word recognition with young learners

As we have seen, the degree of variation in reading speed among young learners varies enormously, and generally, it is recommendable to avoid written word-recognition study using reaction time measurements before 4th grade (9 or 10 years old). In the above case study, a sizeable increase in reaction time occurred with students younger than 4th grade, which was attributed to slow reading speed. When studying with younger learners, it is advisable to use non-RT-based methodologies, such as paper-based testing. While ultimately the methodology will be dictated by the factor(s) under examination in the individual research study, removing reaction time from the equation yields multiple benefits. First, in a standard word recognition scheme (such as lexical decision tasks), slow reading speed can preclude accurate measurement of the actual processing time, as the subject may have made their decision **before** reading is accomplished. For example, in the above case study, it is possible that students, having heard the English word, had already determined the possible Japanese equivalent long before they finished reading the two options. At that point, the RT becomes a measure of reading speed, not of lexical recognition. Using multiple choice paradigm may be an effective means of measuring word recognition in a paper-based format. In such tests, the focus necessarily switches from processing speed to processing accuracy. Also, note that while students at the low end of the allowable range are probably capable of yielding usable RT data, the degree of difference in reading speeds between grades still makes direct cross-grade comparison something to avoid. Differences in latent RTs are more likely due to maturational constraints than from any experimental condition.

4.2 Endeavor to make test-formats "Kid-friendly"

Adapting tests for age-appropriacy is a weighty topic unto itself and demands the attention of many assessment specialists. In the most general terms, it is recommendable to limit the number of items used in any test being conducted with young learners, as this will limit the amount of test-time to a reasonable amount better suited to their limited attention spans. Despite standard psychometric testing protocols which would dictate having hundreds of test items, the researcher must keep in mind that once the child loses interest in the task, the data collected will diminish precipitously in both accuracy and value. As such, it is much better to have fewer data points that are accurate (i.e., that the child paid attention to) than many data points, most of which are unreliable. The general rule of thumb in test timing is to keep everything to under ten minutes for young children, although for 4th graders and older, 20+ minutes may be possible. In the above case study, the test was restricted to 20 items, plus instructions and a practice period. This caused the total testing time to range from 6-10 minutes, depending upon the subject. While this sort of shortened test is certainly more appropriate for young learners, in order to attain reliable results, there are ways of increasing the amount of data collection. For example, when possible, testing could be collected over multiple periods, but this will depend upon the test paradigm. Additionally, if you are testing over long periods of time, be sure to factor in maturational effects. If the researcher is limited in the number of data points collected with each child, the natural means of overcoming statistical limitations is to increase the number of subjects: i.e., whereas one may test 30 adult subjects with 300 test items each, it may be more fruitful to test 300 child subjects with a mere 30 test items. This was the course taken by the researchers in the case study above. While the case study is taken from a limited sample pool used while calibrating the test instrument, for the full study, the researchers increased the number of testing sites (ultimately testing over 1000 children) to counterbalance the relatively low number of test items. Increasing the size of the subject pool is also effective in overcoming the elevated error rates typically seen among child subjects compared to adults.

4.3 Analyze data by age groups

Generally, it is best to analyze individual age cohorts by themselves until at least the age of 15 when, according to the study by Bucsuházy and Samela (2017), child reaction times become functionally identical to that of adults. While there are potential scenarios wherein cross-grade comparisons are the point of the research, still, if the researcher is using reaction time as a central measure of performance, it is crucial to factor maturational effects into the analysis. Slower reaction times among younger learners are generally to be expected. Depending upon data type, however, it can be possible to compare similar or diverging patterns in data while maintaining the assumption that slower overall performance among younger test takers are due to maturational effects and not some other factor (e.g., if 4th graders showed the same overall patterns, but with slower RT than 5th graders, it could be productive to compare the data patterns and to assume the RT differences are maturational in origin). In the above case study, this was the tact taken. While there was initially some suspicion on the part of researchers that we would find two basic groupings with an age "cut-off" wherein children would no longer be susceptible to input from mental concepts to recognizing L2 words, the breakdown of the data immediately revealed that each grade cohort had to be analyzed separately, as their reading speeds varied so significantly as to mostly preclude cross-grade comparison. The older the age cohorts get, the more likely it is they can be productively compared, but researchers should maintain caution throughout the ages of formal education.

Also note that data from very young children (i.e., 3-6-year-olds) may require more limited age groupings (e.g., by months, etc.) as intergroup differences of maturational effects are significantly more profound as one looks at progressively younger ages. Depending upon the focus of testing, the difference between a 3-year-old and a 3 ½ year-old may be highly significant.

5. Discussion and conclusion

Ultimately, while reaction timing (and psychometric techniques, generally) are extremely useful to researchers for investigating issues such as how language is processed and used in real-time, we need to hold in mind that, like any other tool, the usefulness of psychometric techniques is limited to specific purposes. In the case of reaction time studies, there is reason to question their relative usefulness for testing children whose inherent physical and psychological limitations may preclude reliable data yield. While there are already a great many factors that can exert extraneous influence on RT data from adult learners (e.g., handedness, visual acuity, age, gender, etc.), these considerations largely pale before the degree of influence that child maturational effects can have on RT, and this, in turn, can make it extremely difficult for the researcher to ascertain any effect from the variables they are trying to measure. Whenever possible, it is highly advisable that researchers attempt to probe children using non-RT-based research techniques, and where the use of reaction time as a central measure is unavoidable in a study on young children, we must make certain to account for maturational factors in the analysis. By following the protocols discussed in this paper, the researcher aspiring to test child subjects will be better prepared to account for the inevitable complexities caused by child maturational and attentional limitations.

References

Bucsuházy, K. & Samela, M. (2017). Case study: Reaction time of children according to age. *Procedia Engineering, 187*, 408-413.

Forster, K.I. & Forster, J.C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavioral Research Methods, Instruments, & Computers*, *35*, 116-124.

Hasbrouck, J. & Tindal, G. (2017). *An update to compiled ORF norms* (Technical Report No. 1702). Eugene, OR, Behavioral Research and Teaching, University of Oregon.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Academic Press.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

Nunan, D. (1992). *Research methods in language learning*. Cambridge University Press.

O'Shea, G. & Bashore, T. R. (2012). The vital role of *The American Journal of Psychology* in the early and continuing history of mental chronometry. *American Journal of Psychology*, *125*(4), 435-448.

Sanders, A. F. (1998). *Elements of human performance: Reaction processes and attention in human skill*. Lawrence Erlbaum Associates.

Welford, A. T. (1980). Choice reaction time: Basic concepts. In A. T. Welford (Ed.), *Reaction Times* (pp. 73-128). Academic Press.

Williams, C. & Naganuma, N. (2020). Image use effects on young learner vocabulary acquisition. *Proceedings of the 8th OPENTESOL International Conference 2020*, 106-116.

Williams, C.H. & Naganuma, N. (2018) Concept mediation by elementary L2 learners. *PanSIG Journal 2017*, 203-210.