IAEA Conference 2018

Using speech-to-text software for the Hong Kong Diploma of Secondary Education Examination: a case study on the implementation of a new accommodation for candidates with specific learning disabilities

CHU Shun-lung, Kenneth
Hong Kong Examinations and Assessment Authority
E-mail: slchu@hkeaa.edu.hk

Abstract

The Hong Kong Examinations and Assessment Authority (HKEAA) is dedicated to providing reliable and equitable examination and assessment services. To uphold the integrity and fairness of public examinations, all candidates, including those with special educational needs (SEN), are assessed using the same standard. Nonetheless, the HKEAA provides adaptations and accommodation to enable SEN candidates to demonstrate the full extent of their learning and be equitably assessed under suitable conditions without having an unfair advantage over other candidates.

In response to requests from stakeholders to allow candidates with Specific Learning Disabilities (SLD) to dictate their answers in public examinations, mainly due to their functional limitations on handwriting, a working group comprising educational/clinical psychologists and experts in the field of special education was established in 2012/13 to study the feasibility and appropriateness of allowing them to use speech-to-text software for answering essay-type questions orally in the Hong Kong Diploma of Secondary Education Examination (HKDSE)[1].

Subsequent to literature review on benchmarking international practices, field testing of software and user experience tests, an experimental pilot study consisting of training and testing sessions and semi-structured interviews for both SLD and control groups was conducted in 2014/15. In view of the conclusions of the study and recommendation of the working group, the HKEAA has allowed SLD candidates who have severe difficulties in writing and fulfilled the prescribed eligibility criteria to apply for using speech-to-text software for answering questions in the Liberal Studies examination starting from the 2017 HKDSE.

This paper discusses how different data collected and analysed in the aforementioned pilot study support the implementation of the new accommodation for SLD candidates in the HKDSE, and how statistical analysis on the examination performance of the candidates using the software in the 2017 and 2018 HKDSE will be conducted to review the effectiveness of the accommodation.

[1] Normally, the HKDSE is taken by students after three years of senior secondary education.

Background

Over the years, there have been persistent requests from relevant stakeholders to allow SLD candidates to dictate their answers in public examinations, mainly due to their functional limitations on handwriting. With the introduction of the HKDSE in 2012, a working group comprising educational/clinical psychologists and experts in the field of special education was established to study the feasibility and appropriateness of allowing SLD candidates to use speech-to-text software for answering essay-type questions orally in the public examinations. A research team consisting of HKEAA colleagues (including an in-house educational psychologist) was also formed to conduct the literature review, fielding testing and user experience tests of the software, and the experimental pilot study.

Literature Review

In general, the number of research studies concerning the impact of accommodation with speech-to-text software was limited and the results were mixed. For example, one study had proved that both dictation to a scribe and dictation using software helped high school students with learning difficulties (LD) produce higher quality essays while the quality of writing was not improved for students without LD (MacArthur & Cavalier, 2004). With accommodations, students wrote with fewer errors, better quality in terms of ideas, content, organisation, word choice, sentence fluency and conventions. Nevertheless, Tindal & Fuchs (1999) opined that much of the research done on speech-to-text tools reflected post-hoc evaluations with weak internal validity, thus failing to offer conclusive evidence either in support or in criticism of the tool as an accommodation.

The international practice of speech-to-text tools was benchmarked. In the United States, speech-recognition systems are one of the assistive technologies available for students with LD (Day & Edwards, 1996). In Canada, occupational therapists recommend students with handwriting problems complete all or part of their work using dictation strategies but no consensus has been reached regarding the underlying evidence base for deciding technology recommendations for students experiencing handwriting problems (Freeman, MacKinnon & Miller, 2004). In Taiwan, according to the Technology-Related Assistance for Individuals with Disabilities Act (1988), students with disabilities are eligible for assistive technology including speech recognition software. As regards the criteria for identifying students in need of speech-to-text tools, reference was made to the Access Arrangements, Reasonable Adjustments and Special Consideration provided by the Joint Council for Qualifications (JCQ) of the United Kingdom.

The working group was of the view that the research studies were not conclusive in this regard. Some research findings are available to support the use of speech-to-text software for SLD candidates but the level of such evidence is moderately low and there is a need for well-controlled

research in this area. As such, fairness to all candidates should be carefully considered if SLD candidates are allowed to use speech-to-text software in the local public examinations. On one hand, it should provide SLD candidates with an equal opportunity but not an unfair advantage over other candidates; on the other hand, assistance with writing using the tool should increase the likelihood that the test score is a better indicator of what the candidate has learnt in a particular subject. Empirical evidence on this improvement of test validity should be obtained and documented. Several issues were also raised regarding the provision of software, i.e. (i) current capabilities of speech recognition system (e.g. accuracy of the software); (ii) the need for sufficient training for candidates to use the software; (iii) the need for using the software on a regular basis in schools, and (iv) test validity/constructs (e.g. allowing candidates to use speech-to-text software instead of writing will invalidate test integrity if writing mechanics are one of the target areas of test measurement).

Speech recognition software

It was indispensable to identify a brand of speech-to-text software which would tailor the needs of Cantonese speakers within an acceptable voice recognition accuracy range. According to a preliminary testing of several brands of software in the market, only one supported Cantonese translation but the accuracy rate of 50-80% would be a cause for concern. Nevertheless, the HKEAA sought expert advice from the Department of Electronic Engineering[2] of the Chinese University of Hong Kong (CUHK) and the use of MacBook Air was suggested. It would be a speaker-independent system which enables the user to speak Cantonese at a natural pace and have his/her voice, entered as text, into the word processing document. Further to the field testing of the speech-to-text function of MacBook Air conducted in May 2014 in collaboration with the CUHK, a number of user experience tests were conducted to address two issues: (i) whether the software would attain an acceptable level of accuracy in speech recognition and (ii) if any undesirable assistance by artificial intelligence was available. The testing results of accuracy rate were satisfactory with an average of around 90% (see Table 1).

Table 1 *Testing results of accuracy rate*

Target words/sentences	Examples	Average no. of times of dictation before correct recognition	Overall accuracy rate
252 Chinese words at secondary school level*	見識、挫敗、評價	2.1	99.21%
50 Chinese proverbs/idioms*	一般見識、 耐人尋味、 罄竹難書	1.38	100%
Reading of first three words of a Chinese proverb/idioms (total: 50)**	一般見、 耐人尋、 罄竹難	3.04	60%

[2] The Department possesses extensive experience in researching on speech-to-text tools.

Reading of Chinese proverb/idioms with the last word wrongly spoken but with the same onset/rime**	一般見星、 耐人尋命、 罄竹難豬	1.48 (last word with same onset) 2.6 (last word with same rime)	96% (last word with same onset) 72% (last word with same rime)
20 sentences (同音異字 errors / nonsensical words)*	工作*完畢*後，我才可以離開。 我常用*鉛筆*寫作。	4.43	91.72%
210 sentences with random words**	表示中心收集挑戰果園。	4.05	86.29%
Two LS sample answer scripts (total number of words: 1562)		N/A	88.48%

* *The target word was repeated for a maximum of five times. If speech recognition failed to produce the correct word, it will then be matched with other words (配詞) for a maximum of additional five times.*

** *The words were repeated for a maximum of five times. On a few occasions, the software would automatically generate the last word or correct the wrongly spoken word.*

However, some artificial intelligence and undesirable functions which should be disallowed in public examinations were noted. For example, the software would help to select the appropriate word phrase, among a set of phrases with identical pronunciation, based on the context of the voice input. Some 4-character Chinese idioms would also be generated automatically when only the first 3 characters are read. The word association function (which would bring unfair advantages to candidates in public examinations) is made available to users when the trackpad of the MacBook Air is used, and it cannot be disabled.

The functionality of the MacBook Air was also considered by the subject managers of the HKEAA. Several general principles governing the use of the software were established. First of all, it should not violate the assessment objectives of the subjects/papers or pose any undue advantage to the SLD candidates over other candidates. Secondly, it should not be used in language subjects where the writing input of candidates are to be assessed; it is only applicable to subjects/papers requiring extended writing such as long essay-type questions. Lastly, only candidates with severe writing difficulties should be eligible for the provision.

The findings were carefully considered by the working group. It was agreed that the built-in speech-to-text function of MacBook Air would be suitable for the intended purpose (i.e. allowing SLD candidates with severe writing difficulties to answer essay-type questions in written public examinations orally). However, those undesirable functions that might affect the fairness or appropriateness of the accommodation (e.g. internet access, calculation, non-standard autotext, predictive text, dictionary/translation, spell check, grammar-check, thesaurus, word association functions for dictation/voice recognition and input methods, choice of words with identical pronunciation (e.g. '完畢' vs. '鉛筆') by the user) should be removed or disabled prior to the public examinations. Besides, candidates should be prohibited from using the trackpad for editing as the word association function cannot be disabled but they can still manage to edit the text with the allowable Chinese character input methods.

Pilot Study

Under the HKDSE, most candidates would take four core subjects (i.e. Chinese Language, English Language, Mathematics Compulsory Part and Liberal Studies) of which the results would be considered for admission to local tertiary institutions. According to the Regulations and Assessment Frameworks of the HKDSE, the emphasis of Liberal Studies examinations would be on understanding and assessing the extent to which candidates can demonstrate possession of the appropriate thinking skills learnt in the subject, and candidates are required to attempt data-response questions and extended-response questions in either one language version of Chinese or English. As the provision of the software would undermine the test integrity if the assessment is intended to measure writing achievements, Liberal Studies was selected for testing accommodation in this study as it would not compromise the assessment objectives of the subject and the provision would cater to more SLD candidates if found to be appropriate and feasible.

Participants

Seventy-four students (including 37 SLD and 37 control) attending Secondary 5 of 17 mainstream secondary schools participated in this study from September 2014 to March 2015. All the SLD participants must have a confirmed diagnosis of SLD and evidenced severe writing difficulties as measured by the relevant assessment tool for SLD with local norms. No co-morbidity with other types of SENs was reported by the schools. Control group participants were recruited from the same school without history of SEN as reported by the schools. They were also administered relevant assessment tools to ascertain normal intellectual functioning and age-appropriate writing ability respectively. All the participants had to demonstrate proficiency in using the software as their major mode of text production in the testing session.

Procedure

All the participants were given training and practice to use the speech-to-text software in MacBook Air for four one-hour sessions over a four-week period. Apart from teaching participants how to use the software, emphasis was placed on the use of written Chinese for oral dictation (書面語) and use of graphic organisers (e.g. mind maps) to brainstorm and connect concepts in order to assist oral production.

For the testing session, the participants were randomly assigned into the two experimental/treatment groups. They were required to answer two open-ended essay questions in Liberal Studies in the same order (i.e. Part A first and then followed by Part B). Order of answering modes was counterbalanced, with the first group using the software for Part A and handwriting for Part B, and vice versa for the second group (see Table 2). These two questions were selected by the relevant subject manager of the HKEAA based on the learning level of the target students. They were of

comparable level of difficulty testing student's subject knowledge on the same issue.

Table 2 *Grouping of the testing session*

Grouping / Modes of text production	SLD Group		Control Group	
	Group 1	Group 2	Group 1	Group 2
Speech-to-text software	Part A (N=19)	Part B (N=18)	Part A (N=19)	Part B (N=18)
Handwriting	Part B (N=19)	Part A (N=18)	Part B (N=19)	Part A (N=18)
	Total SLD Group (N=37)		Total Control Group (N=37)	

The participants were given the standard time of 75 minutes to finish the two questions but for those who did not manage to finish within the time limit, the standard 25% extra time allowance (ETA) was given and their work done during the ETA was differentiated. The answer scripts, with or without ETA, were then marked separately. The handwritten scripts were typed on a word processor, preserving all errors, while the dictated scripts on MacBook Air were printed, before marking. Two markers were assigned to mark each script independently (i.e. double marking, as the usual marking practice for Liberal Studies in the HKDSE), and they were not told of the participants' modes of text production.

A semi-structured interview was conducted after the testing session to collect students' feedback on the dictation tool (e.g. their preferred mode of text production and why; whether handwriting or speech-to-text software helped them write better and why). Besides, teachers' impression scores on students' test performance against their daily performance were collected.

The handwritten scripts were typed on a word processor, preserving all errors, while the dictated scripts on MacBook Air were printed, before marking. Two markers were assigned to mark the participants' scripts independently (i.e. double marking, as the usual marking practice for Liberal Studies in public examinations). The markers received the scripts in printed copies and they were not told about the participants' answering modes.

Data Analysis

Anaylsis of variance (ANOVA), repeated measures ANOVA, correlation and descriptive analyses were used to interpret the data obtained based on (i) the performance scores given by the HKEAA markers, (ii) measures (e.g. accuracy rate, number of words dictated, speed of production, etc.) rated by the research team, and (iii) feedback provided by the participants, in consultation with the Assessment Technology and Research Division of the HKEAA.

Results and Discussion

This study investigated the appropriateness of using speech-to-text software as examination accommodation for candidates with SLD. The first question addressed by the study was whether the participants would learn to use the software as a mode of producing texts with acceptable accuracy. In general, it appeared that the accuracy rate of speech recognition improved with longer and more coherent phrases. In the pilot study, the average accuracy rates of both SLD and Control groups for all the tasks in the training and testing sessions were over 90% (see Table 3). Both groups did not differ significantly in the number of words dictated using the software in the training sessions, but the SLD group wrote significantly fewer words than Control group in handwriting mode in the testing session (see Table 4). Within group comparison was made to see if there existed any difference between handwriting and software modes but no significant differences were found.

Table 3 *Accuracy rate for all the tasks in the training and testing sessions*

Tasks	SLD (mean)	Control (mean)	p-value	Significant difference between groups?
Reading task (Session 1)	94.4	96.4	0.06	No
Short questions (Session 1)	92.4	95.2	0.19	No
Translation exercise (Session 2)	95.4	96.4	0.34	No
Self-introduction (Session 2)	95.9	96.2	0.81	No
Short Liberal Studies question 1 (Session 3)	96.0	96.9	0.40	No
Short Liberal Studies question 2 (Session 3)	94.8	96.2	0.22	No
Liberal Studies question 1 (Session 4)	96.0	97.4	0.12	No
Liberal Studies question 2 (Session 4)	96.0	97.0	0.23	No
Test (software mode)	94.7	97.3	0.11	No
Test (handwriting mode)	96.6	98.0	0.006	SLD had significantly lower accuracy rate than Control Group

Table 4 *Number of words dictated/written between SLD and Control groups in the testing session*

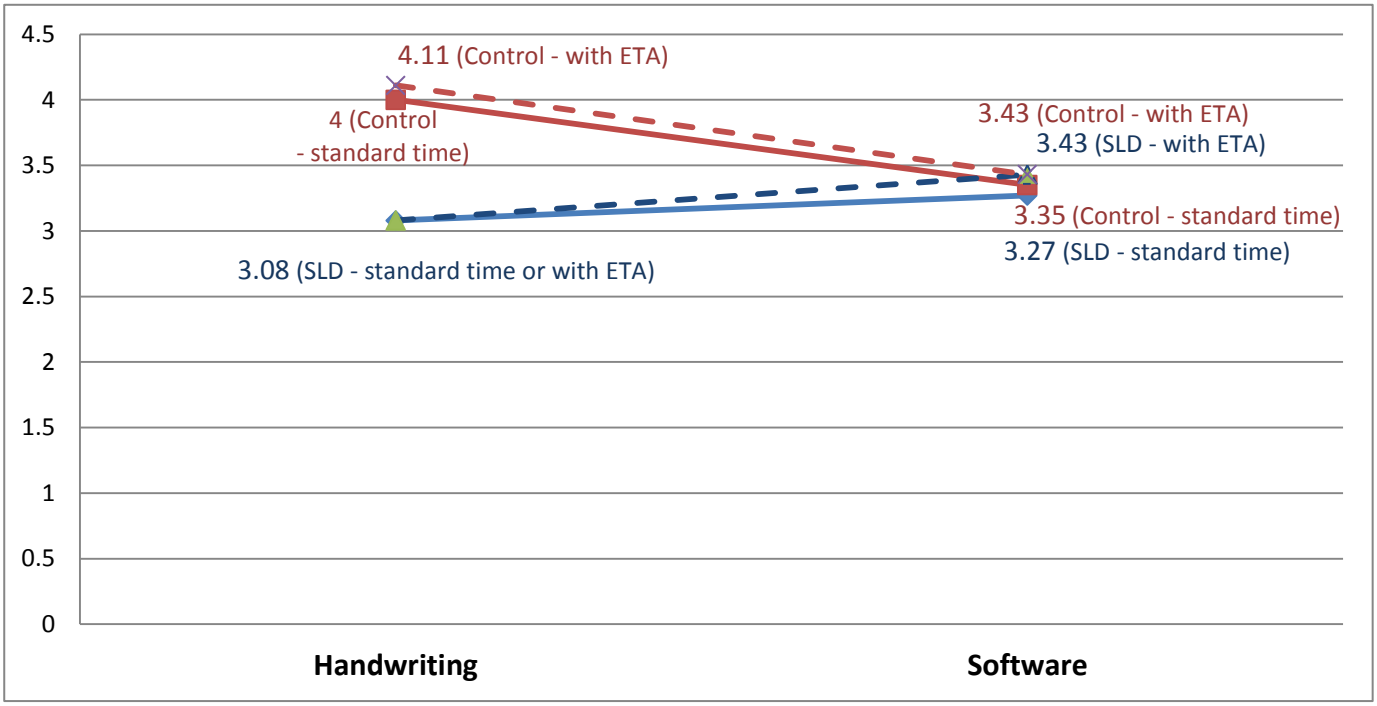
Tasks	SLD (mean)	Control (mean)	p-value	Significant difference between groups?
Test (software mode) – Standard time	270.9	315.7	0.13	No
Test (handwriting mode) – Standard time	270.7	337.8	0.05	Yes
Test (software mode) – With ETA	287.0	329.5	0.21	No
Test (handwriting mode) – With ETA	272.9	356.9	0.03	Yes

The second question addressed was whether the software would help SLD subjects with severe writing difficulties produce better written answers (improved quality of performance) while no statistically significant difference in the quality of performance would be found for the Control group. As illustrated in Table 5, Control group attained higher scores[3] in handwriting than

[3] In line with the Standards-referenced Reporting of assessment results in the HKDSE, the participants' performance was reported against a set of standards divided into five levels (levels 1 to 5), with 5 being the highest. Candidates with the best performance in level 5 are awarded a 5**, and the next top group is awarded a 5*. Attainment below level 1 is designated as "Unclassified".

software mode. Their difference in performance scores (software versus handwriting) was negative and the difference was significant. On the other hand, SLD group performed significantly poorer than Control group in handwriting mode but this difference narrowed down in software mode. Rate of growth was calculated based on the increase in performance scores from handwriting to software mode. SLD group attained 6.2% improvement in standard time and 11.4% with ETA whereas Control group attained -16.3% and -16.5% deterioration in standard time and with ETA respectively. Rate of growth for SLD with ETA was larger than 10%, which could be considered substantial in a real examination setting.

Table 5 *Performance scores between SLD and Control groups*



Words per minute were calculated by dividing the total number of words produced by total time spent on the task. It served as an indicator of how fast or slow the participant worked on a task. As shown in Table 6, Control group dictated significantly fewer words per minute (i.e. "slower") in software mode than handwriting mode but no such difference was noted for SLD group.

Table 6 *Words per minute between software and handwriting modes*

		Software	Handwriting	Difference	p-value	Cohen's d (effect size)	Significant difference between groups?
SLD (N=37)	Mean	10.8	11.9	-1.1	0.20	-0.246	No
	SD	4.0	4.9				
Control (N=37)	Mean	14.3	17.1	-2.8	0.003	-0.474	Yes
	SD	6.1	5.7				

All participants were interviewed after the training and testing sessions. The majority of them

(81% of SLD and 79% of Control group participants) indicated that they found the user experience of the software satisfactory. 84% of the SLD participants reported that they found the software effective in overcoming their barriers in writing. However, only 62% of SLD and 52% of Control group participants shared that the software was easy to use. The major difficulty reported by the participants was accuracy problem of the speech recognition. They claimed that the software often produced words with similar sounds rather than the target words. Extra effort was needed to check if the words were correctly recognized. They also needed to try different ways to improve accuracy such as dictating the target words together with other phrases or use of Chinese character input methods.

Recommendations

The findings were carefully considered by the working group and the HKEAA. There was no undue advantage over non-SLD students as Control group did not benefit from the software. They did poorer in software mode than in handwriting mode. The difference in performance scores (software vs handwriting) was negative and it was statistically significant. They also dictated fewer words per minute in software mode.

SLD group performed significantly poorer than Control group in handwriting mode. They wrote significantly fewer words and attained lower performance scores than Control group. This performance gap narrowed down in software mode for both groups. Most importantly, SLD group had a small increase in the performance scores in software mode, which could be considered as substantial, in terms of rate of improvement (i.e. over 10%) but the magnitude was not statistically significant. It was hypothesised that whether the SLD students could make better use of the software would depend on their knowledge in the subject.

Nevertheless, the software is not 100% accurate and using the software to replace handwriting would create new challenges (e.g. extra effort to plan/organise ideas and edit/modify the text to ensure accuracy). Those who displayed articulation and pronunciation difficulties would face additional problems of poor speech recognition by the software. As such, the software might not be helpful to all eligible SLD candidates and sufficient training should be given to the candidates to get familiar with the software (including the allowable Chinese character input methods) and use of written Chinese for oral dictation. In addition to using it in school internal assessments, the software should be used on a regular basis for academic work. A relatively quiet environment is also required for using the software.

It was concluded that the accommodation of software did not pose any unfair advantage of SLD participants over Control participants; and the magnitude of change (a small increase in performance scores was noted) to SLD participants was not statistically significant, but the rate of improvement could be regarded as substantial. Finally, the new accommodation was implemented

in the 2017 HDKSE for SLD candidates with severe writing difficulties. The fact that the software might not be helpful to all eligible SLD candidates was clearly stated and conveyed to the stakeholders.

Implementation of the New Accommodation

In the 2017 HKDSE, totally 226 candidates with SLD applied to use the software in the Liberal Studies examinations and 220 applications were approved based on their severe writing difficulties as measured by the relevant assessment tool for SLD with local norms. Subsequently, 92 candidates withdrew from using the software after school practice. The majority of the candidates who withdrew opined that they did not manage to use the software well or there was no substantial improvement in performance with the use of the software. As such, only 128 candidates from 67 schools actually used the software in the 2017 HKDSE Liberal Studies examination. A total of 96 special examination venues (including 64 single rooms, 14 standard classrooms, 15 function rooms and 3 school halls) were set up. A survey was administered to the schools and candidates concerned to evaluate the provision of this newly-introduced SEA. Both the schools and candidates concerned were generally satisfied with the overall examination arrangements.

In the 2018 HKDSE, with the upsurge in the number of SLD candidates with severe writing difficulties, 250 out of 256 applications were approved. Subsequently, 155 candidates from 77 schools actually used the software in the Liberal Studies examination. A total of 110 special examination venues (including 60 single rooms, 26 standard classrooms, 22 function rooms and 2 school halls) were set up. When compared to other accommodation for SLD candidates, the withdrawal rate of using speech-to-text software was exceptionally high in the first two years which explained the limitations of the software and the importance of sufficient training in school internal examinations prior to the HKDSE.

Review of the Effectiveness of the Accommodation

To review the effectiveness of the accommodation and whether the provision should be extended to other non-language subjects in the HKDSE, a statistical analysis on the examination performance of the candidates using the software in the 2017 and 2018 HKDSE is underway. The examination results of two assessment components/parts of Chinese Language, which focus on reading and writing skills and are broadly comparable to the assessment objectives of Liberal Studies, would be used for comparison. The target participants are SLD candidates who were granted the use of software in the 2017 and 2018 HKDSE. Their performance (i) within Liberal Studies (consisting of two papers with data-response and extended-response questions respectively) and (ii) between Liberal Studies and the two assessment components/parts of Chinese Language would be compared, both within and between actual users and dropouts of the provision. The standardised scores of the target participants would be used for comparison, in view of the limitation of retrieving their

intellectual profile as a control variable in the statistical study. The age and gender factors would also be controlled. Anaylsis of variance (ANOVA) and t-test would be used to analyse the data.

Conclusion

The provision of speech-to-text software to SLD candidates with severe writing difficulties was found to be fair, appropriate and reasonable in the pilot study as it helps to narrow down the performance gap between SLD and Control group participants though the increase in performance scores of SLD group was not statistically significant. It also supported the rationale that the use of software for SLD candidates should not pose undue advantage over non-SLD candidates as the Control group participants did not benefit from the software. In fact, individual differences exist among SLD candidates and sufficient training should be given to them to familiarise themselves with the software in school internal assessments and daily academic work. Whether the software should be extended to other non-language subjects would require further study and assessment on the effectiveness of the software.

References

Day, S. L., & Edwards, B. J. (1996). Assistive technology for postsecondary students with learning disabilities. Journal of Learning Disabilities, 29, 5, 486-503.

Freeman, A. R., MacKinnon, J. R., & Miller, L. T. (2004). Assistive technology and handwriting problems: what do occupational therapists recommend? Canadian Journal of Occupational Therapy, 71, 3, 150-160.

MacArthur, C. A., & Cavalier, A. (2004). Dictation and speech recognition technology as accommodations in large-scale assessments for students with learning disabilities. Exceptional Children, 71, 43-58.

Tindal, G., & Fuchs, L. (1999). A summary of research on test changes: an empirical basis for defining accommodations. Paper presented at the Mid-South Regional Resource Centre Interdisciplinary Human Development Institute, University of Kentucky and the Office of Special Education Programs, U. S. Department of Education.