National Defence
National Defence Headquarters
Ottawa, Ontario
K1A 0K2

Défense nationale
Quartier général de la Défense nationale
Ottawa (Ontario)
K1A 0K2



# Determining the Appropriateness of Extended Time Accommodations in Standardized Cognitive Ability Testing

Rosaria Furlano
Department of National Defence
Director General Military Personnel Research and Analysis

Shawna Goodrich ; Wendy Darr
Department of National Defence
Director General Military Personnel Research and Analysis

Katherine Gibbard ; Jordan L. Ho
Department of Psychology, University of Guelph

Zhigang Wang
Department of National Deference
Director General Military Personnel Research and Analysis

Melissa D. Pike ; Yannick Provencher; Irene Y. Zhang ;
Emma C. Vreeker-Williamson; Peter A. Hausdorf
Department of Psychology, University of Guelph

Canadian Journal of Behavioural Science
Volume 53, Number 2, 2021, pp. 152–163
(Special Issue: Emerging Research in Industrial-Organizational Psychology in Canada)

Date of Publication from Ext Publisher: April 2021



## Defence Research and Development Canada

**External Literature (P)**
DRDC-RDDC-2022-P052
February 2022



Canada

**IMPORTANT INFORMATIVE STATEMENTS**

This document was reviewed for Controlled Goods by Defence Research and Development Canada using the Schedule to the *Defence Production Act*.

CANADIAN PSYCHOLOGICAL ASSOCIATION
SOCIÉTÉ CANADIENNE DE PSYCHOLOGIE

Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement

ISSN: 0008-400X

2021, Vol. 53, No. 2, 152–163
https://doi.org/10.1037/cbs0000257

SPECIAL ISSUE

# Determining the Appropriateness of Extended Time Accommodations in Standardized Cognitive Ability Testing

Rosaria Furlano[1], Shawna Goodrich[1], Wendy Darr[1], Katherine Gibbard[2], Jordan L. Ho[2], Zhigang Wang[1], Melissa D. Pike[2], Yannick Provencher[2], Irene Y. Zhang[2], Emma C. Vreeker-Williamson[2], and Peter A. Hausdorf[2]

[1] Directorate of Military Personnel Research and Analysis, National Defence, Ottawa, Ontario, Canada
[2] Department of Psychology, University of Guelph

In selection contexts that employ standardized testing, the use of accommodations must be accompanied by evidence to support their appropriateness. Three studies examined the appropriateness of extended time accommodations in cognitive ability testing. Study 1 examined the speededness of the Canadian Forces Aptitude Test (CFAT) using data from 12,555 applicants to establish that speed is not a factor in the CFAT. Study 2 examined the impact of extra time in the completion of a paper-based administration of the practice CFAT for test-takers with and without a learning disability (LD). Data from 122 military trainees revealed that regardless of LD status, participants received higher scores and attempted more items. Study 3 replicated and extended the findings of study 2 using a university sample ($N = 234$) and a computerized test administration mode. Findings from both studies also suggested that construct- and criterion-related validity were comparable across time conditions. Overall, this research has implications for those considering time accommodations in cognitive ability testing, and contributes to the limited body of knowledge on test accommodations available to selection test developers and providers.

***Public Significance Statement***
This study found that providing extra time during standardized test completion benefits both individuals with and without a diagnosed learning disability (LD). Findings suggest that organizations and testing providers should re-examine practices that allow for the provision of extra time to some groups to ensure that they do not unintentionally disadvantage others.

*Keywords:* time accommodations, cognitive ability, standardized testing

*Supplemental materials:* https://doi.org/10.1037/cbs0000257.supp

Human Resources (HR) practices such as recruitment and selection must remain responsive to societal changes in order to maintain their relevance and effectiveness. For example, organizations have long regarded workforce diversity initiatives as being important to gain a competitive advantage in increasingly global and demographically diverse contexts (McCuiston et al., 2004). Regardless, organizations must also remain responsive to legislative requirements (e.g., human rights, employment equity) to ensure their HR practices remain legally compliant. Of relevance to this article is Canada's *Accessible Canada Act* to ensure a barrier-free Canada by the year 2040, through the identification and removal of existing barriers, as well as the prevention of new ones in employment (Minister of Justice, 2020). The Act is meant to benefit all persons, but is especially directed at persons with disabilities, where disability refers to any impairment that may be physical, mental, cognitive, communicative, intellectual, learning, or sensory in nature. In an employment context, this Act re-emphasizes an organization's obligations to ensure fair and equitable employment practices, although organizations are also reminded of their duty to accommodate all persons to improve accessibility to programs and services (Canadian Human Rights Commission [CHRC], n.d.). In an employment context, such accommodations pertain to alternate arrangements or procedures in hiring practices (e.g., selection testing), promotional opportunities, training, as well as workplace adjustments.

Despite the well-intentioned objectives of legislation directed at providing accommodations, organizations face challenges in the practical implementation of programs and procedures to improve accessibility. For example, Jones (1997) discussed issues related to

Rosaria Furlano https://orcid.org/0000-0003-4467-8844
Shawna Goodrich https://orcid.org/0000-0003-3638-300X
Jordan L. Ho https://orcid.org/0000-0002-0216-395X
Melissa D. Pike https://orcid.org/0000-0003-2208-9029
Yannick Provencher https://orcid.org/0000-0002-4607-4844
Irene Y. Zhang https://orcid.org/0000-0002-0853-8505
Peter A. Hausdorf https://orcid.org/0000-0002-3657-4044
Correspondence concerning this article should be addressed to Wendy Darr, Department of National Defence, Directorate of Military Personnel Research and Analysis, 101 Colonel By Drive, Ottawa, ON K1A 0K2, Canada. Email: wendy.darr@forces.gc.ca

the provision of accommodations as per the American Disabilities Act of 1990 (ADA) to improve access to advancement opportunities for persons with disabilities. Jones (1997) identified factors at the individual (e.g., stereotypes, stigma) and organizational (e.g., lack of role models) level as preventing the successful implementation of an HR strategy aligned with the ADA. In their examination of workforce diversity, Ernst Kossek et al. (2003) also highlighted workforce diversity strategies may not manifest as intended at lower levels (team or group level) for various reasons, including the lack of agreement about the importance of diversity.

Within personnel selection, the introduction of alternate procedures or arrangements in standardized testing presents challenges for an organization's test providers or testing agencies. According to the *Standards for Educational and Psychological Testing*, referred to as the *Standards* henceforth, accommodations are defined as "minor changes to the presentation and/or format of the test, test administration, or response procedures that maintain the original construct and result in scores comparable to those on the original test." (American Educational Research Association [AERA], 2014, pp. 58–59). Although the provision of accommodations is supported by the *Principles for the Validation and Use of Personnel Selection Procedures*, referred to as the *Principles* henceforth (Society for Industrial and Organizational Psychology, 2018), there is a stipulated need for documented evidence showing that the construct(s) measured by the test does not change, and that comparable inferences can be made from test scores. Such evidence is especially important when tests are used in high-stakes selection contexts, and when norm-referenced comparisons are made between applicants on the basis of the test (Lovett, 2010; Lovett & Lewandowski, 2015). The *Principles* acknowledge that such empirical evidence may not be feasible to produce, but as Lovett and Lewandowski (2015) suggest, most institutions simply resort to granting accommodations because it is quicker and easier to advocate for an accommodation than to gather the necessary evidence to support its appropriateness.

Requests for extended or extra time in the completion of tests are one of the most frequently requested accommodations by individuals with learning disabilities (LDs; Lovett, 2010). Despite an increase in such requests in postsecondary institutions over the years, there is limited information to guide test providers and testing agencies in making decisions about the appropriateness of such accommodations (Lovett & Lewandowski, 2015; Ofiesh et al., 2005). In addition, much of the extant research on test accommodations has been primarily conducted to inform educational institutions. While there is literature discussing accommodations in the employment context (e.g., Pitoniak & Royer, 2001), to the best of these authors' knowledge, there is no published empirical research that directly informs on accommodations in employment testing contexts. In light of an employer's duty to accommodate, particularly in the context of the *Accessibility Canada Act*, this paper describes research aimed at gathering evidence to determine the appropriateness of extended time accommodations in standardized cognitive ability testing used in military selection. Three studies were undertaken. Study 1 focused on the determination of the speededness of a cognitive ability measure. Studies 2 and 3 examined the impact of providing extra time to test-takers with and without a diagnosed learning disability (LD), as well as the validity of test score inferences under standardized and extended time conditions.

## Study 1: Examining Test Speededness of a Cognitive Ability Measure

The purpose of extended time accommodations is to remove construct-irrelevant variance to enable people with time- and processing-related disabilities to finish an equal proportion of the test as those without disabilities (Mandinach et al., 2005). However, it is important to establish that accommodations intended to eliminate such barriers do not inadvertently reduce the test's capacity to measure the intended ability or aptitude (Lovett, 2010). Cognitive ability tests usually require people to complete numerous cognitive tasks that require suitable processing of mental information, and this processing is a major determinant of whether such tasks are performed successfully (Carroll, 1993). Thurstone (1937) alluded to the hybrid nature of cognitive test items as comprising aspects of both speed and ability. Historical perspectives and research on cognitive ability elucidate salient aspects of the relation between cognitive ability and cognitive speed, which have implications for time accommodations in cognitive ability testing.

### Theoretical Basis: Cognitive Ability and Speed

Examinations of the relationships between general cognitive abilities and cognitive speed have been, and continue to be, pivotal to understanding intelligence (Roberts & Stankov, 1999). Although the relationship is widely acknowledged, the composition and magnitude of the relationship has been disputed for a century (Roberts & Stankov, 1999). Proponents of Spearman's theory suggest that the faster an individual is able to process information, the less time is needed to store information in working memory (Vernon, 1983). Those who adhere to Spearman's theory argue that individual differences in intelligence are partly due to variability in the speed with which operations are executed. Thorndike (1927) suggested that general ability be analyzed as three components that include level, range, and speed, where level refers to the number of correct responses, and range refers to the number of tasks at a given difficulty level. Speed provides an indication of the fluency with which an individual performs reasoning tasks. Recent evidence (see Carroll, 1993) for the three-stratum theory of intelligence suggests that abilities can be classified at three levels of generality: general, broad, and narrow. Among the eight broad second-order ability factors, is a broad speed factor with up to 20 narrow first-order factors associated with cognitive speed.

### Test Speededness

Many tests are a possible combination of cognitive speed and ability, and may exist along a speed-ability continuum. For tests at the speed end of the continuum, individual score differences depend entirely on or to a greater extent on the speed of performance (Ofiesh et al., 2005). For tests at the power end of the continuum, performance differences are not based on speed, and test items tend to increase in difficulty. With both moderately and highly speeded tests, response time variations have a greater influence on accuracy. Test developers have only recently begun to specify the composition of a test's level of power/ability and speed, but timed administrations are also a matter of practical limitations (Lu & Sireci, 2007). The omission of information about a test's speededness has raised serious concerns within the psychometric community. Concerns

pertain to tests being designed to measure speed, without attributing variance from test scores to speed (Lu & Sireci, 2007). Of equal concern is that tests designed to measure ability do not account for variance due to speed. Thus, examining the speededness of a cognitive test is a first step in determining the appropriateness of time accommodations.

### Purpose

The objective of this study was to determine the extent to which speed is a factor in the Canadian Forces Aptitude Test (CFAT), a measure of cognitive ability used in the selection of recruits to the Canadian Armed Forces (CAF).

## Method

### Sample and Measure

The sample consisted of 12,555 CAF applicants who took the English computer-based version of the CFAT in 2017 in a proctored group testing session. The CFAT is a 60-item, multiple choice, time-limited test (60 min with instructions) that includes three subscales: verbal skills (VS; 15 items, 5 min), spatial ability (SA; 15 items, 10 min), and problem-solving (PS; 30 items, 30 min). The test is static in nature; items are presented in the same sequential order to all test takers. Test-takers have the option to skip items and return to them if time permits. The use of aids such as a dictionary or calculator is not permitted during test completion. The CFAT has adequate reliability ($\alpha = .90$; Carter et al., 2012), predictive validity ($\rho = 0.58$; Darr & Saindon, 2018), and utility (Catano, 2017).

### Data Analysis

Methods used in Estrada et al. (2017) were used to separate the speed and power components of this cognitive ability measure to determine its degree of speededness. First, speed quotients (*SQ*) were computed for each item to examine whether items on the CFAT are substantially affected by speededness (Stafford, 1971). The *SQ* decomposes variance for the incorrectly answered items into the proportion of items that were not reached (*U*), and the proportion due to errors (i.e., incorrectly answered items) or omissions (*W*), with the formula $SQi = Ui/(Ui + Wi)$. The *SQi* value indicates the proportion of individuals who did not reach an item among those who did not answer an item correctly or omitted the item. In the present analysis, all items after the last item responded to by each applicant were registered as unreached. The items that were considered omitted were all non-answered items prior to the unreached items; these are items that examinees had read, but chose not to answer. Omitted items were corroborated by response time data which provided an indication of whether the item was viewed. Incorrect items were coded as errors. Items substantially affected by speededness are expected to have values $\geq .30$ (Estrada et al., 2017).

Confirmatory Factor Analysis (CFA) was then conducted using Mplus 8 (Muthén & Muthén, 1998–2017) to estimate a baseline model without a speed factor. The baseline model (Model 1) comprised three correlated factors (see Figure S1) representing the three subtests (VS, SA, PS). This model was compared to a second model (Model 2), which included a speed factor (see Figure S2). A speed factor was specified for each subscale, resulting in a total of six factors with correlations freely estimated among the ability factors and among the speed factors, but fixed to zero between cognitive ability and speed. While the ability factor accounts for differences in speed that are related to cognitive ability, the speed factor represents speed that is unrelated to ability. Thus, Model 2 was designed to separate and estimate the variability in the scores due to ability and speed components. Fit indices—including Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI)—were used to assess goodness of fit between the hypothesized model and the data.

## Results

All *SQi* values are $\leq .23$, which is below the 0.30 criterion recommended by Estrada et al. (2017) for items affected by speededness (see Table S1). However, results indicate a trend, with a gradual increase in *SQi* values for the last items of each of the subtests. To be certain that speed is not a factor in the CFAT, a model was tested with only the last item from each subtest allowed to load on their respective speed factors. Model 1 had excellent fit indices (RMSEA = .03, 90% CI [.029–.030]; CFI = .937; TLI = .935). As expected with a large sample size, the value for the $\chi^2 = 20{,}490$ ($df = 1{,}710$, $p = .00$) was significant. Fit indices indicate that a correlated-three factor model fits the data well. Fit indices for Model 2 with variance due to speed (captured by the last item loading on a speed factor for each subtest) were identical to that obtained for Model 1, indicating that this model does not improve fit (RMSEA = .03; CFI = .937; TLI = .935). Overall, the *SQi* values and CFA results suggest that speed is not a factor in the CFAT.

## Discussion

Using various criteria and analyses, our findings on the CFAT's speededness converged on the conclusion that the CFAT is predominantly a measure of ability or power. These findings are a necessary first step in considerations of extended time accommodations in cognitive testing. Such evidence is especially important in the assessment of cognitive ability, as the theoretical literature on cognitive ability suggests that the rate at which tasks are performed is integral to the construct of intelligence (Roberts & Stankov, 1999).

## Study 2: Impact of Extended Time and Learning Disability on Cognitive Ability Testing

Extended time accommodations typically consist of providing test-takers with either 50% or 100% extra time to complete a test (Lewandowski, Cohen, et al., 2013). The extended time accommodation is thought to improve test access for individuals with LDs who may have attention and processing difficulties by reducing the added cognitive demands associated with time limits (Lewandowski, Cohen, et al., 2013). LDs represent a broad category of significant difficulties that are related to academic achievement (Cortiella & Horowitz, 2014), and are categorized into three sub-types based on the most common academic areas that are affected by the disorder (American Psychological Association [APA], 2013): dysgraphia (difficulties in writing), dyscalculia (difficulties with math), and dyslexia. Although distinct from LDs, Attention-Deficit/Hyperactivity Disorder (ADHD) is characterized by attention problems and hyperactivity (APA, 2013),

and can also affect academic achievement scores and grades (LeFever et al., 2002).

Extended time accommodations can have a positive impact on test scores for people with LDs (Rogers et al., 2016), and can also reduce anxiety for students with ADHD (Lewandowski, Gathje, et al., 2013). Despite research showing the benefits of extended time provisions for test-takers with LDs, there has been some skepticism about the fairness of this practice, because some studies have shown that individuals without LDs also benefit from having extra time to complete tests (Sireci et al., 2005). To ensure that accommodations are appropriate, Phillips (1994) suggested the consideration of five key questions; these questions highlight issues/factors that can be broadly categorized as pertaining to the test itself or the LD (Lovett, 2010). This study is focused on considerations that pertain to the test, which are further described below.

## Differential Benefits of Accommodations

An accommodation is valid or appropriate if it benefits *only* those test-takers who are disadvantaged under standardized testing conditions (Lovett, 2010; Lovett & Lewandowski, 2015). Extended time accommodations intended for test-takers with LDs should benefit only those with LDs; improvement in test scores under extended time conditions should only be observed for this group. Empirically, this effect is referred to as the maximum potential hypothesis (Zuriff, 2000). The differential boost hypothesis (i.e., when LD and non-LD test-taker scores improve under extended time conditions, but larger score improvements are observed for LD test-takers) is also sometimes used to support the validity of accommodations (e.g., Sireci et al., 2005). However, Lovett and Lewandowski (2015) do not support this position as it has fairness implications, particularly on standardized tests. For example, in college entrance examinations, if only a few students (i.e., those with LDs) are given extra time, but findings show evidence for a differential boost hypothesis (i.e., non-LD students also benefit), it would be unfair to deprive non-LD students from having extra time on the entrance exam.

In a survey of college students' perceptions, those with and without disabilities indicated that various accommodations (e.g., extended time, breaks) would be beneficial during high-stakes tests, and that *all* students should have access to such accommodations (Lewandowski et al., 2014). Perhaps for this reason, Hollenbeck (2002) emphasizes that an accommodation must be based on need or necessity, and if it serves the purpose of being beneficial, it must be offered to everyone. Therefore, in addition to demonstrating construct equivalence across test-takers under standardized and accommodated conditions, it is important to ensure that the accommodation does not disadvantage those who take the test under standardized conditions. Examining the differential benefits of accommodations requires empirical comparisons of test scores in both LD and non-LD test-takers across accommodated and standardized conditions.

## Validity of Inferences

This second issue pertains to test score comparability across standard and extended time accommodations. Accommodations are intended to remove construct-irrelevant variance resulting from differences across test-takers on skills that are not assessed by the test (Lovett & Lewandowski, 2015). In other words, accommodations are meant to level the playing field by creating equal opportunity for all test-takers; however, they do not always function that way. Lovett and Lewandowski (2015) suggest that accommodations can sometimes increase rather than decrease construct-irrelevant variance, and that testing agencies often provide little empirical evidence to support their use. While *The Standards* (AERA et al., 2014) acknowledge the need for test accommodations to remove construct-irrelevant barriers, they also require evidence that the construct measured by the test does not change with the accommodation.

Hollenbeck's (2002) list of defining criteria for the appropriateness of accommodations includes "unchanged constructs" and "sameness of inferences." The latter refers specifically to the interpretation of test scores. One must be able to make similar inferences across test scores generated under standard and accommodated test conditions (Hansen et al., 2005). Evidence used to demonstrate construct equivalence include comparisons of the internal factor structure of the test under accommodated and standardized conditions, as well as its relationships with other variables and external criteria (Sireci et al., 2008). Lovett and Lewandowski (2015) suggest the use of multiple techniques in investigating score comparability across conditions. Score comparability is especially important when test scores are used to make norm-referenced comparisons across test-takers. According to Lovett (2010, p. 614), "... norm-referenced testing requires the strongest evidence of comparability across standard and accommodated tasks, and in this situation, accommodations should be avoided unless absolutely necessary and supported by evidence of comparability."

## Purpose

In light of the above considerations, this quasi-experimental study sought to determine the appropriateness of extended time accommodations by examining two research questions: (a) How do LD status (LD vs. non-LD) and time (normal vs. extended time) affect performance (i.e., test scores and number of items attempted) on a cognitive ability test?, and (b) Does the validity of the test change across time conditions?

# Method

## Participants

The sample consisted of 122 military participants (non-commissioned members) who were awaiting occupational training. Of these, 24 participants self-reported that they had a current diagnosed LD and/or ADHD (LD group); the remaining 98 participants without a diagnosed LD and/or ADHD formed the non-LD group. The LD group consisted of individuals with the following self-reported diagnoses: dysgraphia (4.17%), dyscalculia (4.17%), dyslexia (12.50%), other LD (4.17%), unsure of LD type (8.33%), ADHD and dysgraphia (4.17%), ADHD and dyslexia (4.17%), Attention-Deficit Disorder (ADD; 20.83%), ADHD (29.17%), and ADD/ADHD (8.33%). In the LD group, six (25%) reported having used an extended time accommodation in prior school or work settings. The LD group consisted of 20.8% women, 8.4% visible minority, and 0% Indigenous Peoples. The non-LD group

consisted of 17.3% women, 25.5% visible minority, and 5.1% Indigenous Peoples.

## Materials

### *Demographics Questionnaire*

The demographics questionnaire inquired about demographic variables, LD diagnosis, and prior accommodations. Responses to questions about LD diagnosis were confidential, and used only to categorize participants into groups (LD and non-LD).

### *Practice Canadian Forces Aptitude Test (PCFAT)*

The PCFAT is a parallel version of the CFAT (see Study 1 for description) that is available on the CAF applicant portal for potential applicants to practice. A paper-and-pencil version was used in this study. Raw scores are added within each sub-scale (VS, SA, PS) to yield sub-scale scores, and can be summed to yield a total score. The internal consistency reliability of the PCFAT for this sample was lower than typical estimates (total score, $\alpha = .77$; VS, $\alpha = .58$; SA, $\alpha = .79$; PS, $\alpha = .55$).

### *Shipley-2*

The Shipley-2 (Shipley et al., 2009) is a self-administered multiple-choice test measuring cognitive ability in individuals aged 7–89 years. The Shipley-2 has good reliability ($\alpha = .92$), and has also been found to correlate strongly with other tests of cognitive ability, such as the Wechsler Adult Intelligence Scale (Shipley et al., 2009). The composite standard score was used to examine the construct validity of the PCFAT.

### *Cognitive Reflection Test*

The Cognitive Reflection Test (CRT) is a six-item test of cognitive reflection, containing mathematical word problems that involve suppressing an inaccurate answer that impulsively comes to mind (Primi et al., 2016). The long version of the CRT was used in the current study as it is appropriate for participants with both low and high levels of cognitive reflection. The CRT displayed adequate reliability ($\alpha = .72$).

### *Syllogistic Reasoning*

This is an eight-item measure of rational thinking and decision-making that does not require numerical ability (Markovits & Nantel, 1989). The items are worded such that the validity of judgement is in conflict with the plausibility of the conclusion. This measure has been found to be correlated with cognitive ability (Stanovich & West, 2000). The reliability of the Syllogistic Reasoning (SR) ($\alpha = .62$) was in line with past reported estimates ($\alpha = .64$; Toplak et al., 2011).

### *Average Grade*

An item, included in the demographic questionnaire, asked participants to report the average grade percentage received in their most recent academic program (high school, college, or university). This measure was used to make inferences about the criterion-related validity of the PCFAT.

### *Self-Evaluation of Performance on Timed Academic Reading*

The Self-Evaluation of Performance on Timed Academic Reading (SEPTAR; Kleinmann, 2005) is a nine-item self-report measure of an individual's perceived need for extra time when reading and taking exams, rated on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*).

## Procedure

Data were collected over two testing sessions of 2 hr each. Participants completed all materials in paper-and-pencil format. The PCFAT was administered in the same way as the official paper version of the CFAT, with participants listening to and reading instructions for each subtest before completing it. A protocol by Lewandowski, Cohen, et al. (2013) was used, where participants were asked to begin each subscale of the PCFAT using a blue pen during the normal time condition, and were asked to switch to a red pen after the normal time had elapsed. Participants were given extra time (half of the normal time) to complete each subtest using the red pen. Therefore, a repeated measures protocol was used such that all participants were exposed to normal and extended time conditions. Once the PCFAT was complete, participants completed the other tests, and were debriefed about the purpose of the study at the end.

# Results

## Differential Benefits of Extended Time

Test performance in the normal time condition was measured by examining correct and attempted items with the blue pen, and performance in the extended time condition was measured by examining items correct and items attempted with the blue and red pen. Number of items attempted refers to the number of items the participant answered, regardless if they were correct or incorrect. Pre-screening analyses found that participants in the non-LD group were more likely to identify as a visible minority compared to those in the LD group. Matched random sample ($n_{LD} = 24$, $n_{non\text{-}LD} = 24$) analyses, with matching on gender, visible minority status, and Indigenous status, revealed no differences in results from that found with the full sample. A-priori power analysis using an expected effect size of $F^2 = 0.33$ indicated that a total sample size of 48 was needed to detect medium-sized effects.

Two mixed-model analyses of variance (ANOVA) were run to determine the effect of group (LD vs. non-LD) and time (normal vs. extended) on the number of items correct and on the number of items attempted on the PCFAT. Results showed a main effect of time on PCFAT scores, $F(1, 119) = 16.08$, $p < .001$, $\eta^2 = .12$ as well as on items attempted, $F(1, 119) = 33.85$, $p < .001$, $\eta^2 = .22$. In other words, participants in the extended time condition not only scored higher on the PCFAT, but also attempted more items on the test compared to those in the normal time condition. There were significant differences in scores and items attempted on the PS subscale and the overall test across time conditions (see Table 1). There was no main effect of LD status on PCFAT scores,

**Table 1**
*PCFAT Performance Across Time and LD Status Conditions*

| | Time (study 2) | | | | |
|---|---|---|---|---|---|
| | Normal | Extended | *t* | *p* | *d* |
| PCFAT VS score | 9.76(2.18) | 9.79(2.17) | 1.35 | .181 | .02 |
| PCFAT SA score | 11.87(2.65) | 12.14(2.90) | 1.83 | .070 | .01 |
| PCFAT PS score | 16.72(3.30) | 17.48(3.37) | 6.94 | <.001 | .03 |
| PCFAT score | 38.35(6.35) | 39.41(6.26) | 5.76 | <.001 | .11 |
| PCFAT VS attempts | 14.78(0.91) | 14.95(0.53) | 2.35 | .021 | .18 |
| PCFAT SA attempts | 14.80(0.98) | 14.90(0.81) | 2.23 | .028 | .03 |
| PCFAT PS attempts | 27.56(3.44) | 29.40(1.93) | 7.75 | <.001 | .38 |
| PCFAT attempts | 57.14(3.96) | 59.25(2.29) | 7.85 | <.001 | .39 |
| | LD status (study 2) | | | | |
| | LD | Non-LD | *t* | *p* | *d* |
| PCFAT VS score | 10.04(1.88) | 9.69(2.25) | 0.70 | .486 | .17 |
| PCFAT SA score | 12.12(2.95) | 11.81(3.35) | 0.48 | .670 | .10 |
| PCFAT PS score | 16.12(3.19) | 16.87(3.32) | −0.99 | .325 | .23 |
| PCFAT score | 38.29(6.01) | 38.37(6.46) | −0.05 | .959 | .01 |
| PCFAT VS attempts | 14.71(1.00) | 14.80(0.90) | −0.42 | .676 | .08 |
| PCFAT SA attempts | 14.75(1.22) | 14.82(0.91) | −0.30 | .767 | .06 |
| PCFAT PS attempts | 27.46(4.08) | 27.58(3.30) | −0.16 | .876 | .03 |
| PCFAT attempts | 56.92(4.58) | 57.19(3.81) | −0.31 | .760 | .06 |
| | Time (study 3) | | | | |
| | Normal | Extended | *t* | *p* | *d* |
| PCFAT VS score | 9.24(2.20) | 9.26(2.21) | 1.34 | .180 | .01 |
| PCFAT SA score | 11.13(2.80) | 11.18(2.77) | 1.93 | .055 | .02 |
| PCFAT PS score | 14.49(4.16) | 17.27(3.84) | 13.45 | <.001 | .69 |
| PCFAT score | 34.86(6.29) | 37.70(6.61) | 13.33 | <.001 | .44 |
| PCFAT VS attempts | 14.98(.16) | 15.00(.07) | 1.34 | .18 | .11 |
| PCFAT SA attempts | 14.93(.65) | 14.98(.33) | 1.93 | .055 | .09 |
| PCFAT PS attempts | 26.73(3.73) | 29.50(1.53) | 13.45 | <.001 | .97 |
| PCFAT attempts | 56.64(3.94) | 59.48(1.68) | 13.33 | <.001 | .94 |
| | LD status (study 3) | | | | |
| | LD | Non-LD | *t* | *p* | *d* |
| PCFAT VS score | 19.48(4.02) | 18.37(4.44) | 1.75 | .081 | .25 |
| PCFAT SA score | 9.04(3.66) | 11.43(2.52) | −4.66 | <.001 | .89 |
| PCFAT PS score | 13.74(3.16) | 16.16(3.65) | −4.65 | <.001 | .67 |
| PCFAT score | 32.52(5.53) | 36.77(6.16) | −4.83 | <.001 | .70 |
| PCFAT VS attempts | 14.96(0.19) | 14.99(0.08) | −1.14 | .261 | .31 |
| PCFAT SA attempts | 15.00(0.00) | 14.95(0.51) | 0.74 | .462 | .11 |
| PCFAT PS attempts | 24.81(1.67) | 28.02(2.44) | 2.31 | .021 | .33 |
| PCFAT attempts | 58.78(1.66) | 57.97(2.64) | 2.20 | .028 | .32 |

*Note.* PCFAT = Practice Canadian Forces Aptitude Test; LD = Learning Disability; VS = Verbal Skills; SA = Spatial Ability; PS = Problem Solving.

$F(1, 119) = .04$, $p = .837$, $\eta^2 < .001$, or on the number of items attempted, $F(1, 119) = .59$, $p = .443$, $\eta^2 = .005$. There was no interaction between LD status and time on PCFAT scores, $F(1, 119) = .89$, $p = .348$, $\eta^2 = .01$, nor on number of PCFAT items attempted, $F(1, 119) = .45$, $p = .505$, $\eta^2 = .004$. These results suggest that participants received a higher score on the PCFAT when given with extended time, regardless of LD status. As such, both LD and non-LD groups benefitted from extra time. Interestingly, there were no significant differences in LD and non-LD participants' self-reported need for extra time.

## Validity of Inferences

To examine whether similar inferences can be made about the construct of interest (cognitive ability as measured by PCFAT) under normal and extended time conditions, correlations were examined between PCFAT scores and the other measures (see Table 2). Positive associations were observed in both the normal and extended time conditions, and controlling for LD status did not statistically change these magnitudes. Using the *cocor* package in *R* statistical environment (Diedenhofen & Musch, 2015), there were no significant differences in associations across the time conditions between the PCFAT and the Shipley-2 composite ($z = -1.11$, $p = .267$, 95% CI [−.08, .02]), CRT total ($z = -.58$, $p = .562$, 95% CI [−.07, .04]), SR total ($z = 1.02$, $p = .307$, 95% CI [−.03, .09]), and average grade ($z = -1.17$, $p = .241$, 95% CI [−.09, .02]). In addition, internal consistency reliability estimates for the PCFAT remained similar across both conditions. These results indicate that psychometric estimates of the PCFAT were comparable across time conditions.

**Table 2**
*Correlations Between PCFAT and Correlates Across Time Conditions*

| | Study 2 | | Study 3 | |
|---|---|---|---|---|
| | PCFAT normal | PCFAT extended | PCFAT normal | PCFAT extended |
| Shipley-2 | .55** | .57** | .53** | .60** |
| CRT total | .56** | .58** | .34** | .38** |
| SR total | .38** | .36** | — | — |
| Average grade | .17 | .20* | .09 | .05 |

*Note.* PCFAT = Practice Canadian Forces Aptitude Test; CRT = Cognitive Reflection Test; SR = Syllogistic Reasoning.
* $p < .05$. ** $p < .01$.

## Discussion

The goal of this study was to examine the appropriateness of providing individuals with and without LDs additional time to complete a cognitive ability test. Results indicated that participants, regardless of LD status, received higher scores and attempted more items on the PCFAT when provided with extended time. Thus, extended time accommodations benefitted both groups of test-takers, and may not be a fair accommodation if it is only granted to those with LDs. With respect to validity of test score inferences, the results demonstrated that the PCFAT displays comparable convergent and criterion-related validity under normal and extended time conditions. Both findings are important in determining the appropriateness of extended time accommodations in standardized testing situations.

## Study 3: Effect of Extended Time, Learning Disability, and Computerized Cognitive Testing

Study 3 extended the contributions of Study 2 by using a computerized mode of test administration. With the increased popularity of computer-based testing by organizations (Derous & De Fruyt, 2016), it is vital to determine if the effects of time accommodations on test performance differ across test modality (e.g., computerized versus paper-pencil). A meta-analysis on the equivalence of computerized and conventional ability tests found a correlation of .97 for power tests and .72 for speed tests, indicating computerization can affect speed tests (Mead & Drasgow, 1993). More recent work has found positive (Preckel & Thiemann, 2003), null (Bayazit & Aşkar, 2012), and negative impacts (Taherbhai et al., 2012) of computerization on test performance. In addition, some evidence suggests that test-takers take longer completing online tests (Bayazit & Aşkar, 2012), and give faster and less accurate responses on computerized tests (Van de Vijver & Harsveld, 1994).

This study also accounted for the potential that participants in Study 2 were primed with knowledge of being given extended time, which may have impacted their self-perceptions—and in turn, impacted their test-taking strategy (DeMarree et al., 2005). Earlier research on priming suggests that behavior can be altered by increasing the accessibility of behavioral representations which can have unintended and passive influences (Bargh et al., 1996). Computerized administration also allowed for an examination of effects at numerous time intervals, whereas past research has mostly focused on accommodations that are 1.5 times or twice the normal time limit (Elliott & Marquart, 2004; Lewandowski, Cohen, et al., 2013).

### Purpose

In addition to replicating the findings of Study 2 using a computerized cognitive test, this study examined the effects of extended time on test performance across various time intervals.

## Method

### Participants

Two hundred and thirty-four students from a Canadian university participated for course credit or cash compensation. Of these, 172 identified as women, 60 identified as men, and 2 chose not to respond. Participants' ages ranged from 17 to 28 years ($M = 18.91$, $SD = 1.63$). Most participants were White/European (63.25%), with the second-most-common ethnicity being South Asian (9.83%). Overall, 27 participants disclosed having a diagnosed LD and/or ADHD (LD group); the remaining 207 participants formed the non-LD group. The LD group comprised individuals with the following diagnoses: dysgraphia (11.11%), other LD (14.81%), unsure of LD type (14.81%), both ADHD and dyslexia (3.70%), ADD (14.81%), ADHD (44.44%), ADD/ADHD (3.70%), and no specification between having ADD and/or ADHD (3.70%). In the LD group, 15 individuals (55.56%) reported having used an extended time accommodation in prior school or work settings. The LD group consisted of 74.07% women, 25.93% visible minority, and 0% Indigenous Peoples. The non-LD group consisted of 17.3% women, 25.5% visible minorities, and 5.1% Indigenous Peoples.

### Materials

The PCFAT ($\alpha = .85$ [total score], .78 [VS], .81 [SA], .77 [PS]), CRT ($\alpha = .66$), Shipley-2 ($\alpha = .96$), the Test Attitude Survey (Arvey et al., 1990; $\alpha = .94$), and a demographic questionnaire similar to that described in Study 2 were used here.

### Procedure

Participants were recruited from a university research participation pool or responded to an email newsletter sent via student accessibility services. They attended a 120-min testing session located in a computer laboratory on campus, which was proctored by at least two of the study authors. Participants were informed that the tests were being completed to assist CAF efforts to maintain the effectiveness of their selection tools. They were also informed that those scoring in the top 10% would be entered to win one of three $100.00 bonus cash prizes. Participants were not allowed to use their phones or access miscellaneous websites during the session. The PCFAT was administered via computer using Qualtrics. Instructions for each subscale were read aloud by the proctor as participants followed the same instructions on the computer screen. Participants were given a total of 1.5 times the normal time limits to complete each subscale (7.5 min [VS], 15 min [SA], and 45 min [PS]), and were instructed to close the survey after the allotted time was up. They were then given a break before resuming the study to complete the Shipley-2 (paper administration) and remaining measures via Qualtrics. At the end of the session, the proctors debriefed participants as to the exact purpose of the study.

## Results

### Differential Benefits of Extended Time

We ran two separate 2 (within-subjects factor: normal time, extended time) × 2 (between-subjects factor: LD group, no LD group) mixed ANOVAs using the *afex* package in *R* (Singmann et al., 2019) to examine the influence on PCFAT items attempted and PCFAT scores. There was a main effect of extended time on number of PCFAT items attempted, $F(1, 232) = 54.04$, $\eta_G^2 = .063$, $p < .0001$, and PCFAT total scores, $F(1, 232) = 54.04$, $\eta_G^2 = .015$, $p < .0001$, with both being higher in the extended time condition (see Table 1). Significant differences were noted only on the PS subscale (see Table 1). While there was no main effect of LD diagnosis on items attempted, $F(1, 232) = 2.42$, $\eta_G^2 = .007$, $p = .121$, there was a significant effect on PCFAT total scores, $F(1,232) = 11.60$, $\eta_G^2 = .045$, $p < .001$; the mean score was higher in the non-LD group. There was no interaction between LD diagnosis and time on items attempted, $F(1, 232) = 2.41$, $\eta_G^2 = .003$, $p = .122$, or on total scores, $F(1, 232) = 2.41$, $\eta_G^2 < .001$, $p = .122$.

### Validity Inferences

Pearson correlations between scores on the PCFAT the CRT and Shipley-2 were calculated with the *stats* package in *R* statistical environment (see Table 2); the same procedures as in Study 2 were used to test for differences across time conditions. As presented in Table 2, all associations are positive and not statistically different across time conditions, with the exception of the Shipley-2 association which was found to differ ($z = 2.75$, $p = .0059$, 95% CI [.02, .13]), but remained equally strong under both conditions. There was a small non-significant, but positive association between PCFAT scores and average grade which did not differ across time conditions ($z = 1.10$, $p = .2702$, 95% CI [−.03, .10]).

We also ran CFAs to confirm the three-factor structure (VS, SA, and PS dimensions) of the PCFAT data by comparing it to a one-factor model across time conditions. We used marginal maximum likelihood estimation—implemented with the *mirt R* package (Chalmers, 2012)—and assessed fit using Hu and Bentler's (1999) two-index presentation strategy. Under both normal and extended time conditions, the three-factor model (normal time: RMSEA = .050, 95% CI [.047, .052], SRMSR = .058; extended time: RMSEA = .018, 95% CI [.008, .025], SRMSR = .058) yielded acceptable fit to the data compared to the one factor model (normal time: RMSEA = .057, 95% CI [.055, .059], SRMSR = .069; extended time: RMSEA = .034, 95% CI [.030, .039], SRMSR = .069). This was also confirmed by chi-square difference tests.

### Exploratory Analyses

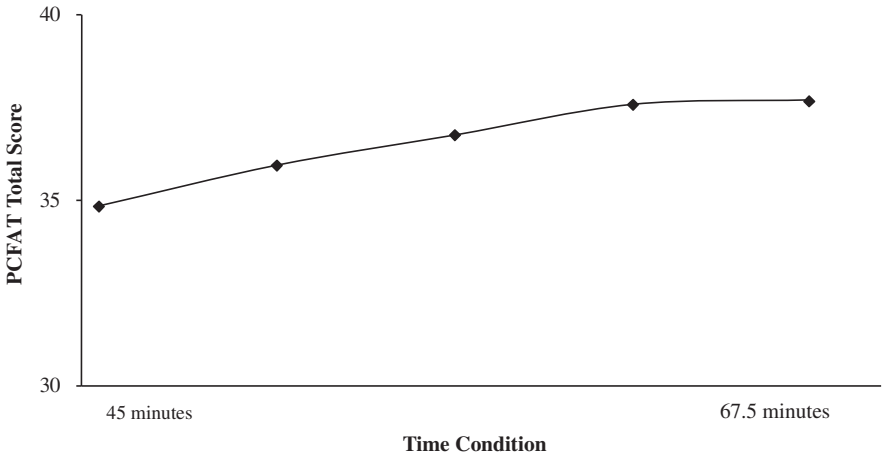
Paired sample *t*-tests comparing PCFAT scores and attempted items across various extended time points (1.375× normal time, 1.25× normal time, 1.125× normal time) showed significantly higher scores and items across all time intervals (see Figure 1, Table S2). Comparisons of subscale scores revealed significant differences for the PS subscale, with $d = .66$, $t(233) = 13.30$, $p < .001$ (1.375× interval); $d = .46$, $t(233) = 13.05$, $p < .001$ (1.25× interval); and $d = .26$, $t(233) = 12.16$, $p < .001$ (1.15× interval). The influence of gender was also explored using mixed ANOVAs (within-subjects factor: normal time, extended time) × 2 (between- subjects factor: male, female) to find no significant main effect for gender or interaction with time on PCFAT outcomes.

## Discussion

This study replicated Study 2 using a computerized mode of test administration. Results were similar to Study 2 in that participants received a higher score and attempted more items on the PCFAT when provided with extended time, regardless of LD status. Hence, these findings do not support the maximum or differential boost hypotheses (e.g., Sireci et al., 2005) that is required for appropriate accommodations. The PCFAT also maintained construct validity under normal and extended time conditions, meeting the criteria of "unchanged constructs" as specified by Hollenbeck (2002). Further, medium to small effects (Cohen, 1988) were found at all time intervals examined (1.375×, 1.25×, and 1.15×) for the total number of items attempted and test scores. However, it appears that extended time only affected performance on the PS subtest.

**Figure 1**

*PCFAT Total Scores Across Time Intervals*



## General Discussion

This paper describes research that must be undertaken by any organization to investigate the appropriateness of an accommodation. According to Standard 3.9 (AERA et al., 2014), test developers and users must furnish evidence to support an accommodation as targeting the test-taker's specific need and as maintaining the construct assessed by the test. Even when an accommodation is not supported, any modifications to the test or testing procedures must be accompanied by evidence that satisfactorily supports its psychometric properties. As illustrated through this examination, a consideration of extended time accommodations in standardized testing must first demonstrate the extent to which a test measures speed. Study 1 illustrated the application of a methodology that empirically established the CFAT as a test of ability or power. Studies 2 and 3 set out to examine the appropriateness of time accommodations by considering two main questions required for this determination. The first pertains to differential benefits; an accommodation is appropriate only if it benefits those that it is intended to benefit (in this case, those with LDs; Lovett, 2010). Results from both studies showed that both groups, LD and non-LD test-takers, benefitted from extra time even when the increases were small (1.125× the normal time limit), suggesting that extended time is not an appropriate accommodation in the administration of this cognitive ability test. These findings furnish important evidence to support decisions about modifications to the test's time limit.

The second question pertains to the validity of test score inferences; validity must be comparable across extended and normal time conditions (Lovett, 2010). Findings from both studies suggest the test's convergent and criterion-related validity were maintained across time conditions. The relevance of grades as a criterion in this research is supported by its use as a key outcome in post-graduate hiring decisions (Strenze, 2007) and its link to job performance ($\rho = .30$; Roth et al., 1996). Moreover, research has demonstrated that GPA is positively related to intelligence test scores (Di Domenico & Fournier, 2015), including cognitive ability measures that are similar to the CFAT such as the Armed Services Vocational Aptitude Battery ($r = .26$, Coyle, 2015). Cognitive ability also reflects people's ability to succeed academically, and academic grades have traditionally been used to establish the criterion-related validity of such tests (Di Domenico & Fournier, 2015). Although our averaged criterion-related validity estimates across time conditions were comparatively lower ($r = .19$, Study 2; $r = .07$, Study 3), these are attributed to possible artifacts (e.g., lower reliability of the PCFAT in Study 2, range restriction in Study 3 due to pre-requisite grade requirements for entry into university) as the meta-analytic validity estimate of the CFAT itself is much stronger ($\rho = 0.58$; Darr & Saindon, 2018). Regardless, the aim of the present research was to demonstrate comparable validities across time conditions, which was supported by our results.

## Limitations

The lower-than-optimal reliability of the PCFAT in Study 2 may be attributable to the paper-based administration of the test, as these estimates were higher in Study 3 which used a computerized administration. We acknowledge that unequal group sizes (i.e., LD and non-LD groups) and unequal variability can impact the power of mean difference tests (Rusticus & Lovato, 2014). However, post-hoc comparisons using a larger hypothetical $n$ of 100 for each LD status group did not change the non-significant findings that were obtained. We note that small samples are typical in developmental psychology research due to the nature of testing and recruiting individuals from specialized populations (Lewandowski, Cohen, et al., 2013).

Another possible limitation is the questionable validity of participants' self-reported LD diagnosis. However, this remains a concern even for professional LD diagnoses. Lovett and Lewandowski (2015) highlighted inconsistencies in the application of criteria in the Diagnostic Statistical Manual-5 (DSM-5) for defining an LD across clinicians and educational institutions, along with problems of subjectivity and measurement error. Research on this topic suggests that many university students with a disability diagnosis fail to meet the standard criteria outlined in the DSM-5 based on test scores (Sparks & Lovett, 2009). Lovett and Lewandowski (2015) acknowledge that there remains a need to advance the science of identifying LD impairments.

## Implications for Research and Practice

Evidence from investigations showcased in this paper are crucial in determining the appropriateness of any accommodation, keeping in mind the definition of an accommodation as "relatively minor changes to the presentation and/or format of the test, test administration, or response procedures . . ." (AERA et al., 2014, p. 58). Organizations have a legal obligation to accommodate individuals with certain needs, particularly if those needs are based on grounds of discrimination (in this case, disability). However, an accommodation may not always be possible because it poses undue hardship for the organization in the form of excessive costs or increased health or safety risks to individuals (Canadian Human Rights Commission [CHRC], 2007). The bona fide justification is often used by organizations who choose to defend an alleged discriminatory practice or the refusal of an accommodation (CHRC, 2007). Successful defences, however, require sufficient evidence (e.g., reason for the existing standard in place, impact of the standard on some groups, consideration of alternate standards, connection of the standard to performance), which is typically obtained through research similar to this one. Therefore, despite legislative requirements, organizations require evidence to legally defend their tools and processes as valid and fair. As Lovett and Lewandowski (2015) suggest, most institutions simply resort to granting accommodations, because it is quicker and easier to advocate for an accommodation than to gather the necessary evidence to support its appropriateness.

While the field of educational research appears to be ahead in addressing issues/concerns pertaining to testing accommodations, there is presently limited guidance available to test developers and administrators responsible for standardized testing in personnel selection contexts. As organizations strive to remain compliant with legislation pertaining to employment accessibility, it is hoped that this paper raises awareness among I/O psychology scientist-practitioners about important considerations in recommending or applying alternate procedures to existing standardized selection practices. As for those involved in selection test/assessment development, the principles of universal design (UD) are especially relevant (Braden & Joyce, 2008; Lovett & Lewandowski, 2015), as they require the anticipation of individual needs and preferences (e.g., preference for visual or auditory stimuli, need for breaks), and

the incorporation of flexibility in the design of a selection test or procedure to address these needs (e.g., shorter tests, large print). The application of UD in test development is likely to result in fewer accommodation requests. It also has the potential to address adverse impact resulting from factors such as language (e.g., Beiser & Gotowiec, 2000) by considering the incorporation of non-verbal test content, for example. In conclusion, this paper adds to the limited body of available guidance to selection testing professionals in making decisions about the appropriateness of accommodations.


## Resumé

Dans les contextes de sélection qui utilisent des tests normalisés, le recours à des mesures d'accommodement doit être accompagné de preuves à l'appui de leur pertinence. Trois études ont examiné la pertinence de mesures d'accommodement de longue durée dans les tests de capacité cognitive. L'étude 1 a examiné la rapidité du Test d'aptitude des Forces canadiennes (TAFC) en utilisant les données de 12 555 candidats pour établir que la rapidité n'est pas un facteur dans le TAFC. L'étude 2 a examiné l'impact du temps supplémentaire dans l'achèvement d'une administration sur papier de la pratique TAFC pour les participants avec et sans trouble d'apprentissage (LD). Les données de 122 stagiaires militaires ont révélé que, peu importe le statut de LD, les participants ont obtenu des scores plus élevés et ont essayé plus d'items. L'étude 3 a reproduit et élargi les résultats de l'étude 2 en utilisant un échantillon universitaire (n = 234) et un mode d'administration des tests informatisé. Les résultats des deux études ont également suggéré que la validité liée à la construction et aux critères était comparable dans le temps. Dans l'ensemble, cette recherche a des implications pour ceux qui envisagent des accommodements de temps dans les tests de capacité cognitive, et contribue à l'ensemble limité de connaissances sur les accommodements de test disponibles pour les développeurs et fournisseurs de tests de sélection.


*Mots-clés :* accommodement de temps, capacité cognitive, tests standardisés

## References


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

American Psychological Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association.

Anon. (n.d.). *Americans with Disabilities Act of 1990 (42 U.S.C. § 12101)*. https://www.eeoc.gov/eeoc/history/35th/thelaw/ada.html

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*, 695–716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.x

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Bayazit, A., & Aşkar, P. (2012). Performance and duration differences between online and paper-pencil tests. *Asia Pacific Education Review*, *13*, 219–226. https://doi.org/10.1007/s12564-011-9190-9

Beiser, M., & Gotowiec, A. (2000). Accounting for native/non-native differences in IQ scores. *Psychology in the Schools*, *37*(3), 237–252. https://doi.org/10.1002/(SICI)1520-6807(200005)37:3<237:AID-PITS4>3.0.CO;2-N

Braden, J. P., & Joyce, L. B. (2008). Best practices in making assessment accommodations. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., pp. 589–603). National Association of School Psychologists.

Canadian Human Rights Commission. (2007). *Bona fide occupational requirements and bona fide justifications under the Canadian Human Rights Act: The implications of Meiorin and Grismer*. https://www.chrc-ccdp.gc.ca/eng/content/bona-fide-occupational-requirements-and-bona-fide-justifications-under-canadian-human-rights

Canadian Human Rights Commission. (n.d.). *What is the duty to accommodate?* https://www.chrc-ccdp.gc.ca/eng/content/what-duty-accommodate

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Carter, J., Boswell, R., Kuschnereit, F., & Ebel-Lam, A. (2012). *Canadian Forces Aptitude Test manual: 2012 update* [Director General Military Personnel Research and Analysis Technical Note 003-2012]. Defence Research and Development Canada.

Catano, V. M. (2017). *Utility analysis of the Canadian Forces Aptitude Test* (Director General Military Personnel and Analysis Contractor Report DRDC-RDDC-2017-C273). Defence Research and Development Canada.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Cortiella, C., & Horowitz, S. H. (2014). *The state of learning disabilities: Facts, trends and emerging issues*. National Centre for Learning Disabilities.

Coyle, T. R. (2015). Relations among general intelligence (*g*), aptitude tests, and GPA. Linear effects dominate. *Intelligence*, *53*, 16–22. https://doi.org/10.1016/j.intell.2015.08.005

Darr, W., & Saindon, M. (2018). *Predictive validity of the Canadian Forces Aptitude Test (CFAT): An updated meta-analysis* (Director General Military Personnel and Analysis Scientific Report DRDC-RDDC-2018-R016). Defence Research and Development Canada.

DeMarree, K. G., Wheeler, S. C., & Petty, R. E. (2005). Priming a new identity: Self-monitoring moderates the effects of nonself primes on self-judgments and behavior. *Journal of Personality and Social Psychology*, *89*, 657–671. https://doi.org/10.1037/0022-3514.89.5.657

Derous, E., & De Fruyt, F. (2016). Developments in recruitment and selection research. *International Journal of Selection and Assessment*, *24*, 1–3. https://doi.org/10.1111/ijsa.12123

Di Domenico, S. I., & Fournier, M. A. (2015). Able, ready, and willing: Examining the additive and interactive effects of intelligence, conscientiousness, and autonomous motivation on undergraduate academic performance. *Learning and Individual Differences*, *40*, 156–162. https://doi.org/10.1016/j.lindif.2015.03.016

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, *10*, Article e0121945. https://doi.org/10.1371/journal.pone.0121945

Elliott, S. N., & Marquart, A. M. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children*, *70*, 349–367. https://doi.org/10.1177/001440290407000306

Ernst Kossek, E., Markel, K. S., & McHugh, P. P. (2003). Increasing diversity as an HRM change strategy. *Journal of Organizational Change Management*, *16*, 328–352. https://doi.org/10.1108/09534810310475550

Estrada, E., Román, F. J., Abad, F. J., & Colom, R. (2017). Separating power and speed components of standardized intelligence measures. *Intelligence*, *61*, 159–168. https://doi.org/10.1016/j.intell.2017.02.002

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System*, *33*, 107–133. https://doi.org/10.1016/j.system.2004.11.002

Hollenbeck, K. (2002). Determining when test alterations are valid accommodations or modifications for large-scale assessment. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 395–425). Erlbaum.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Jones, G. E. (1997). Advancement opportunity issues for persons with disabilities. *Human Resource Management Review*, *7*, 55–76. https://doi.org/10.1016/S1053-4822(97)90005-X

Kleinmann, A. E. (2005). *Not so fast: Using speed to differentiate high and average readers* [Unpublished doctoral dissertation]. Syracuse University.

LeFever, G., Villers, M., Morrow, A., & Vaughn, E. (2002). Parental perceptions of adverse educational outcomes among children diagnosed and treated for ADHD: A call for improved school/provider collaboration. *Psychology in the Schools*, *39*, 63–71. https://doi.org/10.1002/pits.10000

Lewandowski, L., Cohen, J., & Lovett, B. J. (2013). Effects of extended time allotments on reading comprehension performance of college students with and without learning disabilities. *Journal of Psychoeducational Assessment*, *31*, 326–336. https://doi.org/10.1177/0734282912462693

Lewandowski, L., Gathje, R. A., Lovett, B. J., & Gordon, M. (2013). Test-taking skills in college students with and without ADHD. *Journal of Psychoeducational Assessment*, *31*, 41–52. https://doi.org/10.1177/0734282912446304

Lewandowski, L., Lambert, T. L., Lovett, B. J., Panahon, C. J., & Sytsma, M. R. (2014). College students' preferences for test accommodations. *Canadian Journal of School Psychology*, *29*, 116–126. https://doi.org/10.1177/0829573514522116

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, *80*, 611–638. https://doi.org/10.3102/0034654310364063

Lovett, B. J., & Lewandowski, L. J. (2015). *Testing accommodations for students with disabilities: Research-based practice*. American Psychological Association. https://doi.org/10.1037/14468-000

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*, 29–37. https://doi.org/10.1111/j.1745-3992.2007.00106.x

Mandinach, E. B., Bridgeman, B., Cahalan-Laitusis, C., & Trapani, C. (2005). *The impact of extended time on SAT® test performance*. Educational Testing Service (ETS) Research Report Series.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, *17*, 11–17. https://doi.org/10.3758/BF03199552

McCuiston, V. E., Ross Wooldridge, B., & Pierce, C. K. (2004). Leading the diverse workforce: Profit, prospects, and progress. *Leadership & Organization Development Journal*, *25*, 73–92. https://doi.org/10.1108/01437730410512787

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Minister of Justice. (2020). *Accessible Canada Act (S.C. 2019, c. 10)*. https://Laws-lois.justice.gc.ca/eng/acts/A-0.6/

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.).

Ofiesh, N., Mather, N., & Russell, A. (2005). Using speeded cognitive, reading, and academic measures to determine the need for extended test time among university students with learning disabilities. *Journal of Psychoeducational Assessment*, *23*, 35–52. https://doi.org/10.1177/073428290502300103

Phillips, S. E. (1994). High stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, *7*, 93–120. https://doi.org/10.1207/s15324818ame0702_1

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research*, *71*, 53–104. https://doi.org/10.3102/00346543071001053

Preckel, F., & Thiemann, H. (2003). Online versus paper-pencil version of a high potential intelligence test. *Swiss Journal of Psychology*, *62*, 131–138. https://doi.org/10.1024//1421-0185.62.2.131

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*, 453–469. https://doi.org/10.1002/bdm.1883

Roberts, R. D., & Stankov, L. (1999). Individual differences in speed of mental processing and human cognitive abilities: Toward a taxonomic model. *Learning and Individual Differences*, *11*, 1–120. https://doi.org/10.1016/S1041-6080(00)80007-2

Rogers, C. M., Lazarus, S. S., & Thurlow, M. L. (2016, May). *A summary of research on the effects of test accommodations: 2013–2014* (NCEO Report 402). https://nceo.umn.edu/docs/OnlinePubs/Report402/NCEOReport402.pdf

Roth, P. L., BeVier, C. A., Switzer, F. S., III, & Schippmann, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology*, *81*(5), 548–556. https://doi.org/10.1037/0021-9010.81.5.548

Rusticus, S. A., & Lovato, C. Y. (2014). Impact of sample size and variability on the power and type I impact of sample size and variability on the power and type I error rates of equivalence tests: A simulation study. *Practical Assessment, Research, and Evaluation*, *19*, Article 11. http://pareonline.net/getvn.asp?v=19&n=11

Shipley, W. C., Gruber, C. P., Martin, T. A., & Klein, A. M. (2009). *Shipley-2: Manual*. Western Psychological Services.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2019). *Afex: Analysis of factorial experiments*. https://CRAN.R-project.org/package=afex

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment*, *13*, 108–131. https://doi.org/10.1080/10627190802394255

Sireci, S. G., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*, 457–490. https://doi.org/10.3102/00346543075004457

Society for Industrial and Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). Cambridge University Press.

Sparks R. L., & Lovett B. J. (2009). College students with learning disability diagnoses: who are they and how do they perform? *Journal of Learning Disabilities*, *42*(64), 494–510. https://doi.org/10.1177/0022219409338746

Stafford, R. E. (1971). The speededness quotient: A new descriptive statistic for tests. *Journal of Educational Measurement*, *8*, 275–277. https://doi.org/10.1111/j.1745-3984.1971.tb00937.x

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–665. https://doi.org/10.1017/S0140525X00003435

Strenze, T. (2007). Intelligence and socioeconomic success: A meta analytic review of longitudinal research. *Intelligence*, *35*, 401–426. https://doi.org/10.1016/j.intell.2006.09.004

Taherbhai, H., Seo, D., & Bowman, T. (2012). Comparison of paper-pencil and online performances of students with learning disabilities. *British Educational Research Journal*, *38*, 61–74. https://doi.org/10.1080/01411926.2010.526193

Thorndike, E. L. (1927). The measurement of intelligence. Bureau of Publications, Teacher's College, Columbia University.

Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika*, *2*, 249–254. https://doi.org/10.1007/BF02287896

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks.

*Memory & Cognition*, *39*, 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Van de Vijver, F. J., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the general aptitude test battery. *Journal of Applied Psychology*, *79*, 852–859. https://doi.org/10.1037/0021-9010.79.6.852

Vernon, P. A. (1983). Speed of information processing and general intelligence. *Intelligence*, *7*, 53–70. https://doi.org/10.1016/0160-2896(83)90006-5

Zuriff, G. E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education*, *13*, 99–117. https://doi.org/10.1207/s15324818ame1301_5

Received March 17, 2020
Revision received December 11, 2020
Accepted December 14, 2020 ■

## DOCUMENT CONTROL DATA

*Security markings for the title, authors, abstract and keywords must be entered when the document is sensitive

1. ORIGINATOR (Name and address of the organization preparing the document. A DRDC Centre sponsoring a contractor's report, or tasking agency, is entered in Section 8.)

   750 First Street NE, Washington,
   DC 20002-4242

2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.)

   CAN UNCLASSIFIED

2b. CONTROLLED GOODS

   NON-CONTROLLED GOODS
   DMC A

3. TITLE (The document title and sub-title as indicated on the title page.)

   Determining the Appropriateness of Extended Time Accommodations in Standardized Cognitive Ability Testing

4. AUTHORS (Last name, followed by initials – ranks, titles, etc., not to be used)

   Furlano, R.; Goodrich, S.; Darr, W. ; Gibbard, K.; Ho, J.L.; Wang, Z.; Pike, M.D.; Provencher, Y.; Zhang, I.Y; Vreeker-Williamson, E.C.; Hausdorf, P.A

| 5. DATE OF PUBLICATION (Month and year of publication of document.) | 6a. NO. OF PAGES (Total pages, including Annexes, excluding DCD, covering and verso pages.) | 6b. NO. OF REFS (Total references cited.) |
|---|---|---|
| April 2021 | 12 | 68 |

7. DOCUMENT CATEGORY (e.g., Scientific Report, Contract Report, Scientific Letter.)

   External Literature (P)

8. SPONSORING CENTRE (The name and address of the department project office or laboratory sponsoring the research and development.)

   DGMPRA
   Director General Military Personnel Research and Analysis
   NDHQ (Carling), 60 Moodie Drive, Building 9S.2
   Ottawa, Ontario K1A 0K2
   Canada

| 9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.) | 9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.) |
|---|---|
| OTHER | |

| 10a. DRDC PUBLICATION NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.) | 10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.) |
|---|---|
| DRDC-RDDC-2022-P052 | |

11a. FUTURE DISTRIBUTION WITHIN CANADA (Approval for further dissemination of the document. Security classification must also be considered.)

   Public release

11b. FUTURE DISTRIBUTION OUTSIDE CANADA (Approval for further dissemination of the document. Security classification must also be considered.)

   NONE

12. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Use semi-colon as a delimiter.)

time accommodations; cognitive ability; standardized testing

13. ABSTRACT/RÉSUMÉ (When available in the document, the French version of the abstract must be included here.)

In selection contexts that employ standardized testing, the use of accommodations must be accompanied by evidence to support their appropriateness. Three studies examined the appropriateness of extended time accommodations in cognitive ability testing. Study 1 examined the speededness of the Canadian Forces Aptitude Test (CFAT) using data from 12,555 applicants to establish that speed is not a factor in the CFAT. Study 2 examined the impact of extra time in the completion of a paper-based administration of the practice CFAT for test-takers with and without a learning disability (LD). Data from 122 military trainees revealed that regardless of LD status, participants received higher scores and attempted more items. Study 3 replicated and extended the findings of study 2 using a university sample (N = 234) and a computerized test administration mode. Findings from both studies also suggested that construct- and criterion-related validity were comparable across time conditions. Overall, this research has implications for those considering time accommodations in cognitive ability testing, and contributes to the limited body of knowledge on test accommodations available to selection test developers and providers.