Transfert de tests spatiaux pour adultes à de jeunes adolescents

Transferring spatial tests designed for adults to young adolescents

Sophie Charles – sophie.charles@cyu.fr – https://orcid.org/0000-0002-4499-5842

Cergy Paris Université - Paris

Pour citer cet article : Charles, S. (2024). Transfert de tests spatiaux pour adultes à de jeunes adolescents. *Évaluer. Journal international de recherche en éducation et formation, 10*(1), 29-57. https://doi.org/10.48782/e-jiref-10-1-29

Résumé

Notre intérêt s'étant porté sur la mesure des habiletés spatiales de jeunes collégiens, notre recherche vise à proposer des techniques d'analyse qui permettent d'évaluer la pertinence des tests spatiaux auprès de publics spécifiques. Pour évaluer les habiletés spatiales de jeunes adolescents, il s'agit en premier lieu de régler le problème de l'identification d'outils de mesure qui leur sont appropriés, et à défaut, des conditions auxquelles on peut utiliser des tests conçus pour d'autres publics. Pour ce faire, nous nous appuyons sur une revue de littérature orientée vers la validation des tests psychométriques en général, et celle des tests spatiaux en particulier, pour concevoir des outils et procédés d'analyse des tests spatiaux et de leurs qualités métrologiques. Ces derniers sont testés sur deux échantillons de collégiens français pour investiguer l'adéquation de quatre tests spatiaux auprès de ces publics. Nos résultats indiquent que les techniques d'analyse que nous avons développées permettent une étude heuristique de l'adéquation de tests psychométriques pour un public spécifique.

Mots-clés

Tests psychométriques, tests spatiaux, outils d'analyse

Abstract

With the purpose of assessing young secondary-school pupils' spatial abilities, our research focuses on characterising four spatial tests, with a view to proposing techniques that can be used to assess the relevance of these tests with specific audiences. To assess young adolescents' spatial skills, it is first necessary to resolve the problem of identifying the measurement tools that are appropriate for these subjects, and failing that, the conditions under which, tests designed for subjects of other ages, can be used with young adolescents. To do this, we draw on a literature review focused on the validation of psychometric tests in general, and that of spatial tests in particular, to design tools and processes for analysing spatial tests and their metrological qualities. These were tested on two groups of French secondary-school pupils to investigate the suitability of four spatial tests for these subjects. Our results indicate that our analysis tools allow for a heuristic study of the suitability of psychometric tests for specific audiences.

Keywords

Psychometric tests, spatial tests, analysis tools

1. Introduction

Plusieurs études internationales (Agbanglanon, 2019; Branoff et Dobelis, 2012; Steinhauer, 2012) ont mis en évidence le lien significatif entre habiletés spatiales et manipulation de modeleurs volumiques pour des étudiants engagés dans des filières à contenus technologiques. Ces outils de conception assistée par ordinateur étaient au cœur du projet e-FRAN EXAPP_3D, qui visait à « *entretenir et accentuer l'intérêt des élèves du secondaire dans les filières techniques et professionnelles dédiées principalement à la conception et la définition de produits industriels en vue d'améliorer leur réussite scolaire* » (ISAE-Supméca, 2016, p. 2). En l'absence de consensus méthodologique concernant les élèves du secondaire, nous avons dans un premier temps conçu un protocole expérimental comportant des mesures des habiletés spatiales à partir de tests psychométriques standardisés, la collecte d'informations relevant des caractéristiques individuelles, telle la performance scolaire, et des observations de l'activité de modélisation pour un public d'étudiants ingénieurs français en première année d'études (Charles, 2023). Nos premiers résultats ayant confirmé la relation significative entre performance spatiale et performance en Sciences, Technologie, Ingénierie et Mathématiques (STIM) établie dans la littérature (Shea *et al.*, 2001; Wai *et al.*, 2009) pour notre public (Charles *et al.*, 2019), nous nous intéressons à présent à transposer ce protocole au public cible des collégiens visé par le projet EXAPP_3D. Par cet article, nous proposons à la communauté scientifique le processus d'élaboration méthodologique pour évaluer les compétences spatiales de jeunes adolescents. Il s'agit d'abord d'identifier des outils de mesure appropriés à ces sujets, et à défaut, de la mise en œuvre de procédés visant à adapter des tests conçus pour d'autres publics. Pour ce faire, nous investiguons les travaux portant sur la caractérisation des tests spatiaux (Eliot et Macfarlane Smith, 1983 ; Hoyek *et al.*, 2012 ; Yue, 2004, 2006), sur la validation de tests psychométriques visant la mesure de l'habileté spatiale auprès d'adolescents (Ramful *et al.*, 2017) et de jeunes adultes (Cohen et Hegarty, 2012), sur l'adéquation de tests spatiaux utilisés avec des adultes (Agbanglanon, 2019 ; Hegarty, 2018), pour des lycéens (Albaret et Aubert, 1996) et des élèves de primaire et de collège (Hoyek *et al.*, 2012) et sur les méthodes de conception d'outils de mesure psychométriques (Bernaud, 2014 ; Cronbach, 1949 ; Hopkins, 1998). Nous en concevons une grille de caractérisation des tests spatiaux et une grille de caractérisation et de validation des tests psychométriques. Ces outils sont ensuite mis en œuvre pour sélectionner des tests spatiaux, puis en évaluer la pertinence pour notre public cible dans deux expérimentations auprès d'élèves de sixième.

2. Revue de littérature

Notre revue de littérature s'organise en trois temps : premièrement, elle s'oriente vers l'investigation de travaux caractérisant les tests psychométriques en général (Bernaud, 2014) et les tests spatiaux en particulier (Eliot et Macfarlane Smith, 1983 ; Hoyek *et al.*, 2012 ; Yue, 2004, 2006), pour qualifier les outils de mesure qui nous intéressent. Le Tableau 7, en annexe, présente de manière synthétique les caractéristiques des tests spatiaux décrits dans cette partie de notre revue de littérature. Dans un deuxième temps, nous nous intéressons aux études portant sur l'adéquation de tests spatiaux pour des adolescents et sur la conception de tests spatiaux destinés à des enfants, adolescents ou jeunes adultes (Albaret et Aubert, 1996 ; Cohen et Hegarty, 2012 ; Hoyek *et al.*, 2012 ; Ramful *et al.*, 2017). Celles-ci nous permettent de repérer les indicateurs et méthodologies mobilisées par leurs auteurs pour valider leurs outils de mesure. Le Tableau 8, en annexe, présente une synthèse de ces indicateurs et des outils décrits par ces auteurs pour les évaluer. Enfin, pour mieux

caractériser les tests spatiaux, nous décidons d'élargir notre revue de littérature aux ouvrages décrivant des méthodes de conception d'outils de mesure psychométriques (Bernaud, 2014; Cronbach, 1949; Hopkins, 1998), visant à aider les évaluateurs à concevoir des outils fiables au travers de différentes étapes de conception et de validation. Ces recherches portent sur les conditions qui permettent d'établir la qualité d'un test, étape préalable à l'investigation de l'adéquation d'un outil de mesure avec un public différent de celui pour lequel il était conçu. Les Figures 1 et 2 illustrent de manière synthétique les processus de validation qui y sont identifiés. Notre état de l'art se limitera aux aspects qui sont décrits en lien avec notre questionnement de recherche, c'est-à-dire aux éléments qui concernent l'adéquation d'un test avec les répondants et les aptitudes visés.

2.1. Caractériser les tests spatiaux

Dans un premier temps, nous investiguons des travaux caractérisant les tests psychométriques en général (Bernaud, 2014) et les tests spatiaux en particulier (Eliot et Macfarlane Smith, 1983; Hoyek *et al.*, 2012; Yue, 2004, 2006), pour qualifier les outils de mesure qui nous intéressent.

2.1.1 Critères de caractérisation des tests psychométriques

Bernaud (2014) distingue les tests psychométriques selon des indicateurs formels : public visé (niveau requis, tranche d'âge visée[1]), format de l'instrument (papier-crayon, matériels, informatisé ou en ligne), mode d'administration (temps limité dans le cas d'un test de vitesse ou libre pour les tests de puissance[2], épreuve individuelle ou collective, consignes, place et rôle de l'évaluateur), mode de réponse (expression orale, écrite ou par geste), mode de traitement des observations recueillies (format de cotation, système d'interprétation) ; ou selon les méthodes employées : tests de performance[3], questionnaires d'autoévaluation, méthodes d'observation, tests implicites. Eliot et Macfarlane Smith (1983) complètent cette liste avec la qualité des instructions verbales (longueur, abstraction, densité) (Eliot, 1983), la raison pour laquelle le test a été conçu (à des fins de sélection professionnelle, d'orientation scolaire ou expérimentales) et la disponibilité du test, c'est-à-dire s'il est commercialisé, épuisé, disponible ou expérimental (Eliot et Macfarlane Smith, 1983).

La modalité de passation s'avère être un critère déterminant : les tests spatiaux papier-crayon sont préférés aux tests de performance, dans lesquels il est demandé au sujet de produire des réponses motrices plutôt que verbales à l'aide de manipulations ou de mouvements physiques (American Psychological Association, s.d.b), car ils sont plus simples à administrer (Lohman *et al.*, 1987). Récemment, on remarque une tendance à l'informatisation des tests papier-crayon spatiaux les plus populaires (Charles, 2023). Cette dernière modalité permet une gestion optimisée de la correction, mais aussi de recueillir des

[1] On remarque cependant que certains tests sont utilisés dans des recherches auprès de publics d'âge variable (Albaret et Aubert, 1996; Hoyek *et al.*, 2012; Vandenberg et Kuse, 1978).

[2]« *a type of test intended to calculate the participant's level of mastery of a particular topic under conditions of little or no time pressure. The test is designed so that items become progressively more difficult* [un type de test prévu pour calculer le niveau de maîtrise d'un sujet particulier pour des participants dont les conditions de passation sont peu ou pas limitées dans le temps. Le test est conçu de telle manière que les items deviennent progressivement plus difficiles] » (American Psychological Association, s.d.c).

[3]« *any test of ability requiring primarily motor, rather than verbal, responses, such as a test requiring manipulation of different objects or completion of a task that involves physical movement* [tout test d'habilité nécessitant principalement des réponses motrices, plutôt que verbales, tel un test nécessitant la manipulation de différents objets ou la complétion d'une tâche impliquant un mouvement physique] » (American Psychological Association, s.d.b).

informations supplémentaires, tel que le temps de réponse (Branoff, 2000). L'informatisation de tests conçus pour être utilisés sous un format papier soulève cependant la question de leur normalisation, c'est-à-dire de la validité et de la fidélité de tests conçus pour mesurer une ou des compétences selon des conditions d'administration, de cotation et donc d'interprétation des scores, spécifiques (American Psychological Association, s.d.e).

2.1.2 Critères de caractérisation spécifiques aux tests spatiaux

Selon Lohman (1993, p.1482), l'habileté spatiale concerne la capacité à « *générer, retenir, récupérer et transformer des images visuelles bien structurées [generate, retain, retrieve, and transform well-structured visual images]* ». Tartre (1990) distingue quatre habiletés spatiales : la rotation mentale, la transformation mentale, le changement de perspective et la capacité à dissocier un élément intriqué dans un motif complexe. Elles peuvent être mesurées au travers de tests psychométriques, dont il existe un nombre important : le répertoire de tests spatiaux de référence d'Eliot et Macfarlane Smith recense 392 tests spatiaux papier-crayon en 1983. Selon Eliot (1983a), ces outils ont d'abord été conçus dans le but d'établir l'existence d'un facteur spatial qui s'ajouterait au facteur général de l'intelligence, puis de définir plusieurs sous-facteurs spatiaux, avant d'établir le lien qu'ils entretiennent avec d'autres habiletés et performances. Ils sont alors utilisés à des fins de sélection et de recrutement aux États-Unis d'Amérique à l'approche de la seconde guerre mondiale et à la suite de l'*Education Act* en 1944 au Royaume Uni. Leur caractère prédictif de la réussite en STIM (Shea *et al.*, 2001 ; Wai *et al.*, 2009) encourage leur usage à des fins d'identification d'élèves aux performances spatiales faibles pour leur proposer des dispositifs de remédiation (Sorby *et al.*, 2013 ; Sorby, 2005 ; Veurink *et al.*, 2009).

La caractéristique des tests spatiaux concerne sans doute le fait qu'ils visent à mesurer des compétences non verbales, et que les tâches qui sont présentées dans les tests papier-crayon sont, à notre connaissance, tous de nature visuelle. La publication du répertoire de tests spatiaux de référence d'Eliot et Macfarlane Smith (1983) a nécessité une classification que les auteurs ont centrée sur la nature des tâches présentées, qu'elles relèvent d'une ou de plusieurs catégories (Eliot, 1983b). De manière à aider les chercheurs à choisir le test le plus approprié pour leurs objectifs de recherche parmi ceux présentant des tâches de même nature, ils ont inclus, pour chaque test répertorié, le texte des instructions et les questions d'entraînement. Ils expliquent ce choix en raison de la variété de représentations de stimuli (en couleur ou en noir et blanc, sous forme de photographie ou de dessin, figures en deux ou en trois dimensions), de leur positionnement dans la page d'instruction (taille, netteté, encombrement de la page).

À l'instar d'Eliot et Macfarlane Smith (1983), Hoyek, *et al.* (2010) ont relevé différentes typologies de stimuli utilisés dans des tâches visant à mesurer la rotation mentale : figures abstraites, segments corporels, objets familiers manipulables ou non. Ils notent que le choix de cette modalité a des effets sur la stratégie de résolution adoptée par les sujets (mobilisation de processus moteurs, stratégie globale ou d'étape en étape), ainsi que sur le temps de réponse. Hoyek *et al.* (2012) mettent par ailleurs en évidence une meilleure performance des enfants dans des tâches de rotation mentale présentant des objets en deux dimensions (2D) que celles impliquant des objets en trois dimensions (3D), et suggèrent qu'il leur est plus facile de manipuler des stimuli représentant des objets familiers plutôt que des figures abstraites. De plus, Yue (2006) note une prédominance de solides en perspective axonométrique dans les tests, *i.e.* des solides dont seules trois surfaces qui

présentent un sommet commun sont représentées (Yue, 2004), avec une préférence pour la projection isométrique qui présente les trois surfaces en proportions égales, car ces représentations sont faciles à dessiner (Yue, 2006). Cependant, Yue (2006) souligne que ces représentations omettent différentes caractéristiques d'un objet, tels les effets de l'éclairage et de perspective, produisant des représentations qui manquent de réalisme. Le Tableau 7, en annexe, présente de manière synthétique les caractéristiques des tests spatiaux décrits dans cette partie de notre revue de littérature.

2.2. Caractérisation des études portant sur la conception et l'adéquation de tests spatiaux à de jeunes sujets

Dans un second temps, nous nous intéressons aux études portant sur l'adéquation de tests spatiaux pour des adolescents et sur la conception de tests spatiaux destinés à des enfants, adolescents ou jeunes adultes, et relevons quatre études (Albaret et Aubert, 1996; Cohen et Hegarty, 2012; Hoyek *et al.*, 2012; Ramful *et al.*, 2017). Les travaux d'Albaret et Aubert (1996) et d'Hoyek *et al.* (2012) concernent l'adéquation du *Mental Rotation Test* (MRT) (Vandenberg et Kuse, 1978) avec des enfants et des adolescents. Les recherches de Ramful *et al.* (2016) et Cohen et Hegarty (2007; 2012) concernent la conception de tests spatiaux adaptés à des adolescents dans le cas du *Spatial Reasoning Instrument* (SRI) (Ramful *et al.*, 2017) et à des jeunes adultes en ce qui concerne le *Santa Barbara Solids Test* (SBST) (Cohen et Hegarty, 2007, 2012). Selon les études, nous relevons des critères relevant de la standardisation, de la fidélité, de la validité, de la fiabilité et de la sensibilité d'un test, et des caractéristiques individuelles qui peuvent l'influencer. Le Tableau 8, en annexe, présente une synthèse des indicateurs et des outils décrits par ces auteurs pour les évaluer. Ce premier état de l'art, qui met en évidence des critères d'évaluation divergents, nous invite à investiguer plus avant les méthodes de conception d'outils de mesure psychométriques (Bernaud, 2014; Cronbach, 1949; Hopkins, 1998), pour mieux comprendre comment évaluer les tests spatiaux.

2.3. Établir les propriétés psychométriques d'un test

D'après Le Corff *et al.* (2017), un instrument psychométrique peut être qualifié à partir de trois caractéristiques : la standardisation, l'objectivité de la mesure et les propriétés psychométriques. La standardisation relève de l'uniformisation du mode d'administration et de correction, de manière à pouvoir comparer les performances des sujets quels qu'ils soient (Cronbach, 1949 ; Le Corff *et al.*, 2017). L'objectivité de la mesure concerne aussi bien la méthode d'évaluation, qui doit être la même quel que soit l'évaluateur qui la réalise (Cronbach, 1949 ; Le Corff *et al.*, 2017), que la conception du test ou ses modalités d'administration. Par exemple, le niveau de langage doit être approprié et ne pas empêcher les sujets de répondre, à moins que l'on cherche à mesurer la compétence de lecture (Hopkins, 1998). Un autre aspect concerne le temps de réponse qui doit être approprié : selon Hopkins (1998), 90% des répondants doivent avoir terminé l'épreuve dans le temps imparti. Un pourcentage inférieur indiquerait que la vitesse est un facteur de réussite, et non pas la maîtrise seule de la compétence visée. Les propriétés psychométriques, ou qualités métrologiques (Bernaud, 2014), quant à elles, « renseignent sur la qualité de la mesure fournie par l'instrument, notamment en termes de précision (fidélité, sensibilité et spécificité) et de validité » (Le Corff *et al.*, 2017).

2.3.1 Validité d'un outil de mesure

La validité d'un test est décrite comme étant la propriété psychométrique la plus importante (Bernaud, 2014 ; Cronbach, 1949 ; Le Corff *et al.*, 2017). Elle concerne la capacité d'un test à mesurer ce pour quoi il a été conçu, autrement dit la justesse des inférences faites à partir des performances mesurées (Hopkins, 1998). On distingue principalement trois types de validité : la validité de contenu concerne la cohérence entre les items d'une part, et les contenus et procédés cognitifs visés par le test d'autre part (Bernaud, 2014 ; Hopkins, 1998 ; Le Corff *et al.*, 2017) ; la validité de critère, ou empirique, relève de la capacité d'un test à évaluer le degré d'association entre les résultats mesurés et une variable indépendante, telle la performance scolaire (Hopkins, 1998 ; Le Corff *et al.*, 2017). Autrement dit, elle évalue les liens entre les résultats produits et des indicateurs choisis (Bernaud, 2014). On distingue la validité concourante, lorsque la prise d'information concernant le critère choisi est simultanée avec la mesure, de la validité prédictive qui fait référence à une prise d'information du critère ayant lieu plus tard que la mesure (Bernaud, 2014 ; Hopkins, 1998). Bernaud (2014) alerte sur la nécessité d'appliquer une correction au coefficient de corrélation, indicateur de validité critérielle, tenant compte de la variance due à la sélection des participants. Il précise aussi que les corrélations de valeur prédictive entre tests d'intelligence et réussite scolaire ou professionnelle dépassent très rarement 0,50. Bernaud (2014) et Hopkins (1998) attirent finalement l'attention sur la validité d'apparence, qui concerne la valeur subjective de la capacité du test à évaluer le construit visé que les évalués et les évaluateurs lui confèrent : selon ces auteurs, une faible valeur apparente peut affecter la motivation des répondants et la qualité des réponses.

La validité d'un test peut donc être vérifiée grâce à des méthodes logiques, en vérifiant notamment la conformité des items du test avec les contenus et procédés cognitifs que l'on cherche à évaluer au travers du test, et dans ce cas, la présence de prérequis fortuits doit être vérifiée (Cronbach, 1949 ; Hopkins, 1998 ; Le Corff *et al.*, 2017). Pour ce faire, Cronbach (1949) recommande de réaliser des expérimentations dans lesquelles on demande aux répondants de décrire à voix haute leur stratégie de résolution pour en vérifier l'adéquation avec la compétence visée et l'absence de prérequis. La validité peut être aussi vérifiée grâce à des méthodes empiriques, en comparant les résultats d'un test avec un autre critère, comme la performance à une tâche (Cronbach, 1949 ; Hopkins, 1998 ; Le Corff *et al.*, 2017) ou d'un autre test visant le même construit et dont la validité a été établie (Cronbach, 1949 ; Le Corff *et al.*, 2017).

2.3.2 Fidélité d'un test

Une deuxième propriété psychométrique est la fidélité d'un test, c'est-à-dire la précision (Cronbach, 1949 ; Hopkins, 1998 ; Le Corff *et al.*, 2017) et la constance (Bernaud, 2014) de la mesure. Ceci fait référence à la fois à « la capacité d'un test à produire un score qui est le plus proche possible du score vrai[4] du sujet » et à « l'obtention de résultats hautement similaires lorsqu'une personne est évaluée à l'aide du même instrument psychométrique à deux moments dans le temps » (Le Corff *et al.*, 2017). Trois méthodes sont principalement utilisées pour estimer le coefficient de fidélité d'un test (Le Corff *et al.*, 2017). Ce coefficient montre l'étendue de l'influence des erreurs de mesure (e.g. chance, circonstance défavorable) sur les scores d'un test (Cronbach, 1949). Il est aussi impliqué dans le calcul de

[4] *« in classical test theory, that part of a measurement or score that reflects the actual amount of the attribute possessed by the individual being measured [en Théorie Classique, la part d'une mesure ou d'un score qui reflète la quantité réelle de l'attribut mesuré possédée par l'individu] »* (American Psychological Association, s.d.-a).

l'erreur type de mesure, qui évalue « *le niveau de fluctuation due à l'infidélité de la méthode* » (Bernaud, 2014, p. 96). Selon Bernaud (2014), la valeur d'un coefficient de fidélité est correcte si elle est égale à 0,70, satisfaisante si elle est égale à 0,80 et élevée si elle est égale à 0,90.

La méthode de stabilité dans le temps consiste à vérifier la corrélation entre les scores obtenus à un test administré à deux reprises, la seconde fois ayant lieu après un laps de temps (Cronbach, 1951 ; Hopkins, 1998 ; Le Corff *et al.*, 2017), autrement dit à vérifier la stabilité des scores dans le temps (Bernaud, 2014, p. 96). Il est alors important de prendre en compte la durée de l'intervalle de temps entre les deux passations pour éviter un effet de maturation dans le cas d'un délai long, ou que le sujet ne se souvienne des réponses qu'il a données lors de la passation précédente, dans le cas d'un délai court (Le Corff *et al.*, 2017). Bernaud (2014) distingue le coefficient de stabilité, pour un intervalle de plus de deux mois, du coefficient de confiance, pour un intervalle inférieur à deux mois.

La méthode d'équivalence propose de vérifier la corrélation entre la performance à deux formes parallèles d'un test (Cronbach, 1951 ; Hopkins, 1998 ; Le Corff *et al.*, 2017) administrées successivement (Cronbach, 1949). Cette méthode est cependant chronophage, pose le problème de l'équivalence de contenu et d'écart type des deux versions (Cronbach, 1949), et peut causer une fatigue ou un ennui chez les administrés (Hopkins, 1998 ; Le Corff *et al.*, 2017). Ces mêmes inconvénients apparaissent quand on combine la méthode de stabilité avec la méthode des équivalences, en administrant une forme parallèle d'un test à une seconde prise de performance, après un laps de temps (Le Corff *et al.*, 2017).

La troisième méthode consiste à vérifier la cohérence interne d'un test, soit le degré de corrélation entre ses items (Cortina, 1993). Deux méthodes sont possibles : dans le cas de la méthode de bissection, on vérifie la corrélation entre la moitié des réponses d'un test avec celles de la seconde moitié (Le Corff *et al.*, 2017). On vérifie ainsi la cohérence de la moitié d'un test avec l'autre (Hopkins, 1998 ; Kuder et Richardson, 1937). En choisissant de comparer les résultats des questions paires d'un test à ceux des questions impaires, on évite l'éventuel effet de fatigue sur les dernières questions qui peut affecter le résultat, si on compare les réponses de la première moitié avec la seconde moitié (Hopkins, 1998 ; Le Corff *et al.*, 2017). Bernaud (2014) recommande donc cette pratique pour les tests non limités dans le temps. Pour compléter cette mesure, on peut utiliser la formule de Spearman-Brown (Brown, 1910; Spearman, 1910) pour estimer la cohésion de l'entièreté du test, en s'appuyant sur le coefficient de fiabilité des deux moitiés (Bernaud, 2014 ; Hopkins, 1998). Ceci suppose que les items des deux moitiés soient issus d'un même univers, autrement dit qu'ils mesurent le même facteur (Cortina, 1993 ; Hopkins, 1998).

En revanche, la méthode des covariances s'appuie sur l'affirmation que la fidélité d'un test repose sur le calcul d'un coefficient de fidélité entre les performances à tous les items d'un test (Hopkins, 1998 ; Le Corff *et al.*, 2017). Il peut être calculé avec la formule Kuder-Richardson 20 (Kuder et Richardson, 1937) qui repose sur l'intercorrélation entre les items qui composent un test. Selon ses auteurs, cette méthode est valable pour les items dichotomiques et peut être utilisée pour des tests dont les items varient en difficulté. Cette formule est un cas particulier de l'alpha de Cronbach (1951) qui généralise la formule KR-20 aux variables continues : α est la moyenne de tous les coefficients de bissection possibles. Selon son auteur, il peut être utilisé pour des sous-tests et des items de difficulté variable. Cortina (1993) démontre cependant qu'il peut être influencé par le nombre d'items. Dans le cas d'un nombre important, ce chercheur recommande de consulter

l'estimation de la précision d'α,, pour vérifier que plusieurs facteurs corrélés ne sont pas mesurés, et le coefficient d'intercorrélation entre les items, pour confirmer cette interrelation. Pour les deux méthodes de vérification de la cohérence interne, il faut rappeler que la stabilité du test dans le temps n'est pas vérifiée (Cortina, 1993 ; Le Corff *et al.*, 2017).

Hopkins (1998) indique que la méthode de la bissection et celle des covariances ne sont appropriées que pour les tests de puissance, i.e. les tests étant peu ou non limités dans le temps (American Psychological Association, s.d.c). Dans les tests de vitesse, certaines questions peuvent rester sans réponse, non pas parce que le sujet ne sait pas y répondre, mais parce qu'il n'a pas eu le temps d'y répondre. Dans ce cas, on obtiendrait des coefficients de fidélité faussement élevés. De plus, il existe le risque que le répondant priorise une des deux injonctions, c'est-à-dire répondre correctement ou dans le temps imparti, ce qui pourrait produire un score qui ne relève pas de la compétence mesurée. L'auteur recommande d'administrer deux parties comparables d'un test selon des temps de passation différents et de vérifier leur cohérence avec la méthode de la bissection, la méthode de stabilité ou la méthode d'équivalence. Bernaud (2014, p. 99) attire aussi l'attention sur la nécessité d'utiliser « *des échantillons suffisants et sélectionnés de façon non aléatoire* » et de combiner plusieurs méthodes, e.g. une mesure de la cohérence interne et une mesure de la stabilité.

2.3.3 Sensibilité

Autrement nommée finesse discriminative (Bernaud, 2014), la sensibilité d'un test concerne sa capacité à différencier les répondants. La forme de la distribution des scores donne des indications sur son adéquation (Cronbach, 1949 ; Hopkins, 1998). Un test trop facile, caractérisé par une queue de distribution étalée vers la gauche, distingue peu les différences de performance entre les sujets moyens et les sujets forts, ni celles au sein du groupe des sujets forts. En revanche, un test trop difficile, caractérisé par une queue de distribution étalée vers la droite, distingue peu les performances élevées et peut contenir des réponses devinées ; à la différence d'une distribution normale qui montrerait plus facilement les scores aux deux extrémités. Bernaud (2014, p. 93) considère quant à lui qu'une distribution gaussienne « *constitue […] une nécessité préalable* ». Il recommande aussi d'observer les moyennes et dispersions (étendue, variance ou écart type), et décrit le delta de Ferguson (1949) comme un outil d'évaluation fine du niveau de sensibilité. Ce coefficient s'appuie sur une approche de la capacité d'un test à distinguer les individus en maximisant la probabilité d'observer des différences entre les individus, au moyen de la sélection d'items de difficulté moyenne, ce qui se traduit par une distribution des scores proche d'une distribution rectangulaire (Ferguson, 1949). Ferrando (2012) précise que cet indice de probabilité dépend de la population testée et s'appuie sur des scores bruts. Il recommande donc qu'il soit utilisé comme une mesure de discrimination auxiliaire.

2.4. Synthèse de l'étude des propriétés psychométriques d'un test

La Figure 1 présente de manière synthétique les caractéristiques des tests psychométriques, ainsi que leurs outils de validation, que nous avons relevées dans la littérature, limitées aux éléments pertinents pour le transfert d'un test psychométrique à un autre public que celui pour lequel il a été conçu.

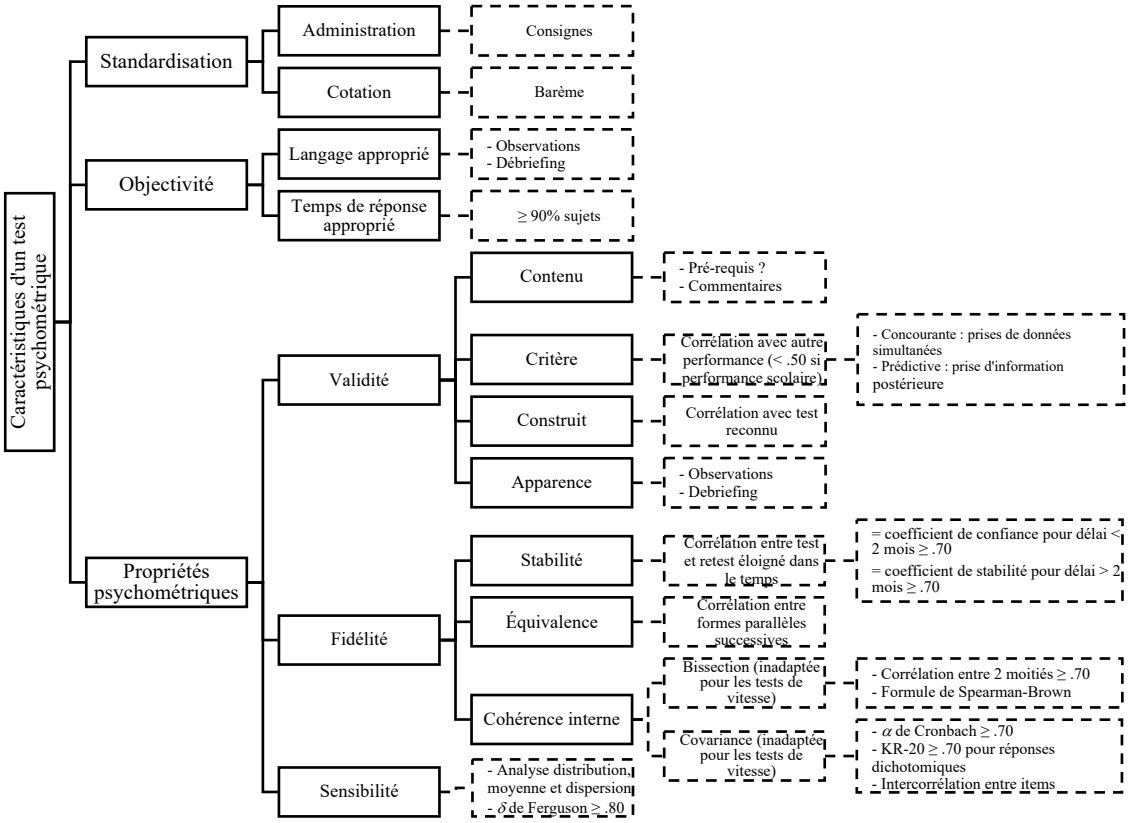

Figure 1 : Synthèse des caractéristiques d'un test psychométrique, accompagnées de leurs outils de validation[5]

2.5. Autres indicateurs de qualité

Bernaud (2014) décrit les étapes, illustrées dans la Figure 2, qu'il juge nécessaires à la validation des items d'un test : dans un premier temps, il préconise, auprès d'un échantillon conséquent (*i.e.* plusieurs centaines de sujets), une pré-expérimentation pour tester le temps de réponse et les consignes et observer les « *réactions et remarques spontanées exprimées par les répondants au cours de l'administration* » (p. 80), recueillir leurs remarques posttest et leur perception de la difficulté des items. Il s'agit ensuite d'expérimenter le test standardisé auprès d'un autre échantillon conséquent, dont les caractéristiques démographiques (*e.g.* âge, niveau de formation, sexe) seront proches de celui de la validation de l'échelle. À l'issue de ce recueil de données, il recommande de calculer l'indice de difficulté, soit le ratio nombre de réponses correctes/nombre total d'items. Bernaud considère qu'un ratio compris entre 0,10 et 0,90 est acceptable. Il attire de plus l'attention sur les éventuels biais culturels dans le cas d'une adaptation d'une épreuve étrangère aux répondants : altération de la signification, signification du construit mesuré selon la culture, taille de l'échantillon, faible familiarité de l'item dans le nouvel échantillon. Il recommande une traduction en retour dans le cas d'une adaptation dans une autre langue.

[5] Les outils d'évaluation des caractéristiques apparaissent encadrés en pointillés.

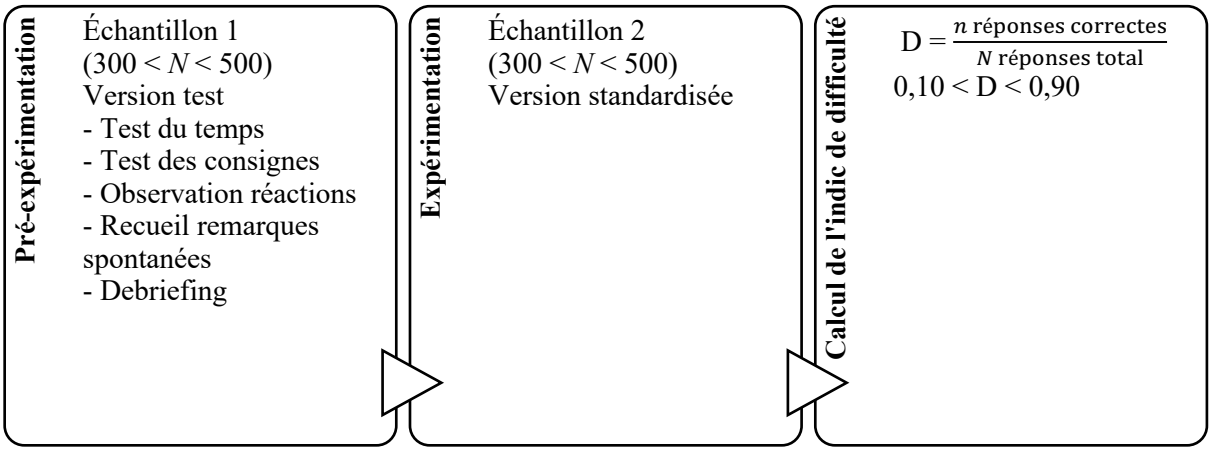

Figure 2 : Procédé de validation des items de Bernaud (2014)

3. Méthodologie

Notre recherche vise la mise à l'échelle du protocole expérimental de mesure des habiletés spatiales que nous avions déployé auprès d'étudiants ingénieurs (Charles, 2023). Ce dernier avait mobilisé une batterie de tests spatiaux visant à mesurer les quatre capacités spatiales de la classification de Tarte (1990) : le MRT et le *Revised Purdue Spatial Visualization Test*: *Rotations* (R PSVT:R) (Yoon, 2011) visent à mesurer la rotation mentale ; le *Mental Cutting Test* (MCT) (*College Entrance Examination Board*, 1939) a pour objet d'évaluer la transformation mentale ; le *Purdue Spatial Visualization Test: Visualization of Views* (PSVT:V) (Guay, 1976) vise la capacité à reconnaitre un objet à partir de plusieurs points de vue ; le *Closure Flexibility Test (Concealed Figures) Form A* (CFT) (Thurstone et Jeffrey, 1965) sollicite la capacité à dissocier un élément intriqué dans un motif complexe.

Nous nous inspirons de la méthode de validation des items de Bernaud (2014), illustrée dans la Figure 2. Nous présentons ci-après la pré-expérimentation, l'expérimentation et le calcul de l'indice de difficulté. La pré-expérimentation comprend un pré-test des habiletés spatiales, une séance de modélisation 3D et un posttest dans un collège du nord de la France, auprès de quatre classes de sixième en mai 2022 (N = 90 ; N_F = 44 filles et N_G = 46 garçons) et en juin (N = 81 ; N_F = 41 filles et N_G = 40 garçons). Les classes ont été choisies par la direction du collège en fonction des disponibilités des enseignants ayant accepté de participer à nos expérimentations. Suite à ces premiers retours, nous modifions la batterie de tests spatiaux et la pratique de la modélisation pour une expérimentation dans un collège parisien en octobre 2022 (N = 100 ; N_F = 53 filles et N_G = 47 garçons), et en mars 2023 (N = 95 ; N_F = 50 filles et N_G = 4 garçons) auprès de quatre classes choisies par l'enseignant participant au projet EXAPP_3D.

Dans les deux situations expérimentales, un formulaire de consentement a été présenté au préalable aux parents expliquant les objectifs et la démarche de la recherche. Nous décrivons ici les données des élèves dont les parents ont donné leur accord. Nous retenons les données des élèves ayant participé à la totalité du protocole et ayant respecté les consignes (*e.g.* donner deux réponses par item pour le MRT) lors de la passation des tests spatiaux, de manière à évaluer leur performance spatiale plutôt que leur respect des consignes. Pour tous les tests spatiaux qui ont été utilisés, nous avons demandé aux élèves de ne pas répondre, plutôt que de répondre au hasard, et nous leur avons expliqué que les scores n'étaient pas pris en compte dans l'évaluation de leur performance scolaire. En

accord avec notre problématique, nous ne présentons ici que les résultats concernant la mesure des habiletés spatiales.

3.1. Pré-expérimentation

Nous décrivons ci-dessous la phase de pré-expérimentation que nous avons réalisée dans un collège du nord de la France.

3.1.1 Méthodologie

Afin de mesurer les quatre capacités spatiales de la classification de Tarte (1990) chez de jeunes adolescents, nous optons pour le SRI, le SBST et le CFT. Ces tests sont choisis pour plusieurs raisons : le SRI est conçu pour prendre en compte les concepts spatiaux couverts dans les curriculums de mathématiques chez des élèves âgés de 11 à 13 ans (Ramful *et al.*, 2017) et présente des questions contextualisées, relevant des quatre facteurs spatiaux de Tartre (1990) et dont les instructions changent à chaque item. Le SBST vise à mesurer la transformation mentale, comme le MCT que nous avions utilisé avec les étudiants ingénieurs, mais il présente des formes moins complexes, et la présence de couleurs distinctes pourrait faciliter la compréhension des représentations. Nous conservons le CFT, que nous avions utilisé avec les étudiants ingénieurs, car il doit sa difficulté au nombre élevé de problèmes à résoudre en un temps très limité, mais les tâches en elles-mêmes ne présentent pas de difficulté particulière. Nous utilisons la version française du SRI de Pierre Chastenay de l'Université du Québec à Montréal, sur les recommandations d'un des auteurs du test (Ramful, communication personnelle, 30 mars 2019). Nous modifions certaines phrases pour les adapter au langage de la cible. Nous utilisons notre version française du SBST, que nous avons traduit avec la permission d'une des auteures, et du CFT, que nous avions utilisé avec des étudiants ingénieurs (Charles, 2023). Nous avons lu les instructions avec les élèves et répondu à leurs questions. Nous obtenons de plus les moyennes scolaires des deux semestres des élèves auprès de la direction pour investiguer la validité critérielle des tests spatiaux, au travers de l'étude des liens entre scores spatiaux et performance scolaire. Les moyennes des deux semestres sont fusionnées pour obtenir une moyenne annuelle en mathématiques d'une part, et en sciences et technologie d'autre part.

De manière à vérifier la pertinence de cette batterie pour notre public cible, nous l'avions soumise auparavant à un échantillon restreint, composé de deux filles et de deux garçons âgés de 12 ans : nous n'avions pas observé de difficulté concernant les instructions lors de la passation. Nous avions relevé les temps de réponses du SRI et du SBST : les statistiques descriptives, décrites dans le Tableau 1, nous ont incité à réduire la durée initiale du SRI de 45 minutes à 20 minutes et celle du test de puissance SBST à 10 minutes, de manière à positionner les trois tests dans l'heure allouée par le collège. Nous maintenons la durée de 10 minutes du CFT qui est un test de vitesse pour lequel il n'est pas attendu que les sujets répondent à l'ensemble des questions dans le temps imparti (Thurstone et Jeffrey, 1956).

Tableau 1 : Statistiques descriptives des temps de réponse, exprimés en minutes et en secondes, du SRI et du SBST pour l'échantillon restreint (N = 4)

Variable	*M*	*ET*	Min	Max
SRI	11:19	03:30	08:30	16:00
SBST	07:24	02:15	05:27	10:00

Note. *M* = moyenne ; *ET* = écart type.

3.1.2 Résultats

Nous présentons ici les résultats obtenus à partir des données relevées en mai 2022, c'est-à-dire la première passation de la pré-expérimentation, pour les élèves présents aux deux recueils (N = 80 ; N_F = 40 filles et N_G = 40 garçons).

Nous contrôlons la normalité des distributions pour les moyennes de mathématiques et de sciences et technologie et les scores spatiaux au moyen d'un test de Shapiro-Wilk (Shapiro et Wilk, 1965) dans le logiciel SPSS (version 28.0). Nous constatons que les moyennes de mathématiques et de sciences et technologie ne suivent pas la loi normale (Tableau 2), avec un étalement relativement important des notes faibles et une concentration des notes de mathématiques entre 13 et 18, et des notes de sciences et technologie entre 13 et 17 (Figure 3).

Tableau 2 : Test de normalité de Shapiro-Wilk des scores des tests spatiaux SRI, SBST et CFT, des moyennes annuelles de français, mathématiques et sciences et technologie (N = 80)

Variable	Statistique de test	*dl*	*p*
SRI - Nord	0,97	80	0,086
SBST - Nord	0,97	80	0,063
CFT - Nord	0,97	80	0,096
Français – Nord	0,94	80	< 0,01
Mathématiques - Nord	0,93	80	< 0,01
Sciences et technologie - Nord	0,96	80	< 0,05

Note. *dl* = degré de liberté ; *p* = valeur de p.

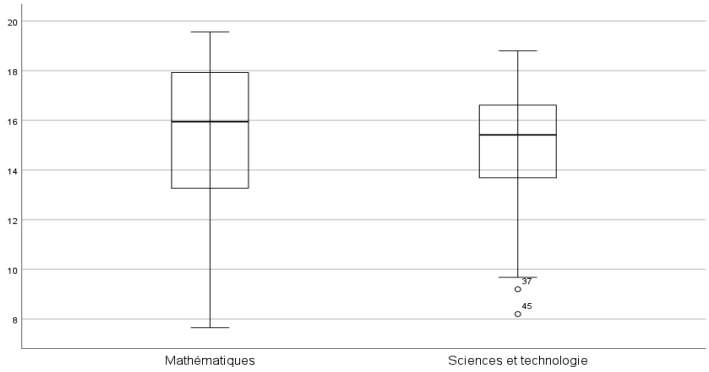

Figure 3 : Distribution des moyennes annuelles de mathématiques et de sciences et technologie des élèves nordistes (N = 80)

Nous reprenons, dans un premier temps, les critères de validation des items de Bernaud (2014) illustrés dans la Figure 2. Nous remarquons, durant la passation du SRI, que trois élèves posent des questions relatives à la compréhension des instructions. De manière à vérifier un éventuel biais du niveau de français des élèves sur la performance au SRI, nous calculons une corrélation par rangs de Spearman (1910), car la distribution de la moyenne de français n'est pas normale (Tableau 2). Nous observons une corrélation très significative ($p < 0{,}001$) et un coefficient de corrélation ($\rho = 0{,}58$) supérieur à la valeur prédictive d'un test psychométrique de la performance scolaire habituellement observée (Bernaud, 2014). Nous calculons les mêmes coefficients pour le SBST et le MRT pour vérifier que cette corrélation n'est pas spécifique au SRI. Bien que les corrélations soient significatives, les coefficients sont inférieurs à celui du SRI. Ces résultats sont repris dans le Tableau 3.

Tableau 3 : Statistiques descriptives et corrélation par rangs de Spearman pour les scores spatiaux et la moyenne annuelle de français ($N = 80$)

Variable	*M*	*ET*	1	2	3	4
1. SRI - Nord	17,74	5,64	—			
2. SBST - Nord	13,34	4,82	—	—		
3. CFT - Nord	30,81	11,28	—	—	—	
4. Français - Nord	14,80	3,07	0,58***	0,27*	0,49***	—

Note. *M* = moyenne ; *ET* = écart type.
* $p < 0{,}05$; ** $p < 0{,}01$; *** $p < 0{,}001$

Nous confirmons ces statistiques par l'observation des nuages de points, dont nous constatons que la distribution des scores du SRI montre que ces scores sont regroupés autour de la droite de corrélation (Figure 4A), à la différence des scores du SBST et, dans une moindre mesure, du CFT (Figures 4A et B). Nous en concluons que la relation entre performance au SRI et niveau de français est trop significative pour écarter la possibilité que le test ne mesure pas seulement l'habileté spatiale, mais aussi la capacité des élèves à lire et à comprendre des instructions écrites.

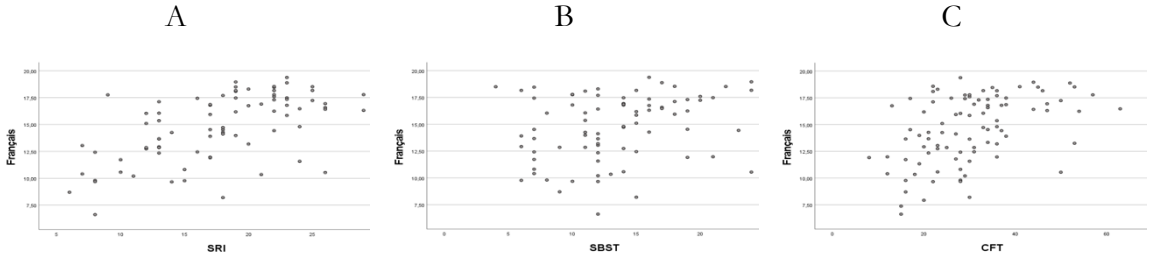

Figure 4 : Nuages de points de la dispersion des scores du SRI (A), du SBST (B) et du CFT (C) en fonction de la moyenne annuelle de français

Nous n'observons pas de difficulté concernant les instructions lors de la passation du SBST. Nous notons, lors de la correction des tests, que cinq élèves [5,5 %] n'ont pas fini le SBST dans les 10 minutes que nous avons allouées à ce test (absence de réponse au moins aux cinq derniers items). Bien que ce chiffre soit supérieur à la recommandation d'Hopkins (1998), nous considérons que cinq [16,5 %] questions non répondues représentent une portion importante de cette épreuve qui a eu lieu en fin d'année scolaire, c'est-à-dire à un moment où les élèves sont des sixièmes confirmés. Nous en concluons que notre restriction du temps de réponse à 10 minutes est trop importante pour certains élèves de sixième.

Nous n'observons pas de difficulté concernant les instructions lors de la passation du CFT. C'est un test de vitesse pour lequel il n'est pas attendu que les sujets répondent à l'ensemble des questions dans le temps imparti (Thurstone et Jeffrey, 1956), soit 49 questions en 10 minutes. Nous ne retenons donc pas cet indicateur pour la validation des items.

3.2. Expérimentation

Nous décrivons ci-dessous la phase d'expérimentation que nous avons réalisée dans un collège parisien.

3.2.1 Méthodologie

Nous décidons d'écarter le SRI en raison des difficultés de lecture rencontrées lors de la première expérimentation et d'étendre le temps de passation du SBST à 15 minutes, en raison des cinq [5,5 %] élèves nordistes qui n'ont pas fini le SBST dans les 10 minutes précédemment allouées. Nous utilisons notre version française du SBST et du CFT (Charles, 2023). Nous optons pour le MRT pour mesurer la rotation mentale car son temps d'administration est court, et qu'il est décrit dans la littérature (Hoyek *et al.*, 2012) comme adapté à des collégiens. Nous maintenons son temps de passation à trois minutes par partie et trois minutes de pause, et utilisons la version française d'Albaret et Aubert (1996). Notre étude vise à concevoir une batterie de tests spatiaux visant les quatre compétences du modèle de Tartre (1990) mais le manque de temps ne nous permet pas d'intégrer un test de mesure du changement de perspective. Nous obtenons de plus les notes scolaires des élèves auprès de la direction pour investiguer la validité critérielle des tests spatiaux au travers de l'étude des liens entre scores spatiaux et performance scolaire. Ces enseignements étant regroupés sous la même discipline en sixième (Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche, 2015), nous calculons la moyenne annuelle des notes de sciences de la vie et de la terre, de physique-chimie et de technologie, qui sont enseignées et évaluées séparément dans le collège parisien. Nous calculons aussi la moyenne annuelle de mathématiques. Nous avons lu les instructions avec les élèves et répondu à leurs questions. En raison du temps nécessité par les explications du SBST et du MRT, nous n'avons pas pu utiliser le CFT.

3.2.2 Résultats

Nous présentons ici les résultats obtenus à partir des données relevées à Paris pour les élèves présents aux deux recueils et qui ont respecté les instructions (e.g. deux réponses par item exigées par les consignes du MRT) ($N = 92$; $N_F = 50$ filles et $N_G = 42$ garçons). Les proportions filles-garçons sont proches de celles du collège nordiste. Nous observons empiriquement, lors du recueil d'octobre, une difficulté des élèves à maintenir leur attention lors de la présentation des instructions et de la réalisation des tâches. Ce constat n'est pas renouvelé lors du recueil de mars.

Nous contrôlons la normalité des distributions pour les moyennes de mathématiques et de sciences et technologie et les scores spatiaux. Nous constatons que les moyennes de mathématiques et de sciences et technologie ne suivent pas la loi normale (Tableau 4), avec un étalement important des notes faibles et une concentration des notes de mathématiques entre 11 et 17, et des notes de sciences et technologie entre 11 et 16 (Figure 5). La distribution des scores de MRT est analysée dans la section 3.4.5. Sensibilité.

Tableau 4 : Test de normalité de Shapiro-Wilk des scores des tests spatiaux MRT et SBST, des moyennes annuelles de mathématiques et de sciences et technologie (*N* = 92)

Variable	Statistique de test	*dl*	*p*
MRT – Paris	0,91	92	< 0,001
SBST – Paris	0,97	92	0,06
Mathématiques - Paris	0,93	92	< 0,001
Sciences et technologie - Paris	0,94	92	< 0,001

Note. *dl* = degré de liberté ; *p* = valeur de p.

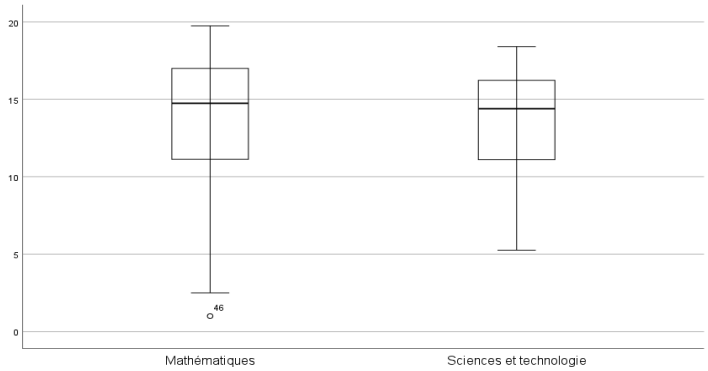

Figure 5 : Distribution des moyennes annuelles de mathématiques et de sciences et technologie des élèves nordistes (*N* = 80)

Nous reprenons les critères de validation des items de Bernaud (2014) illustrés dans la Figure 2. Nous n'observons pas de difficulté concernant la compréhension des instructions lors de la passation du SBST. Nous observons lors de la correction des tests du recueil d'octobre que 8 [8,1 %] élèves n'ont pas fini le SBST dans les 15 minutes que nous avons allouées à ce test, que nous attribuons à une absence de réponse aux cinq derniers items. Ce chiffre diminue pour atteindre un [1,1%] élève lors du recueil de mars.

Nous n'observons pas de difficulté concernant les instructions lors de la passation du MRT. C'est un test de vitesse pour lequel la plupart des sujets adultes ne parviennent pas à répondre à l'ensemble des questions dans le temps imparti (Hegarty, 2018). Nous ne pouvons donc pas retenir cet indicateur pour la validation des items.

3.3. Calcul de l'indice de difficulté

Les trois indices de difficulté sont compris entre 0,10 et 0,90 : $D_{SBST} = 0,43$; $D_{MRT} = 0,20$; $D_{CFT} = 0,16$.

3.4. Analyse des propriétés psychométriques des tests expérimentés

Nous présentons, dans cette partie, l'analyse des trois tests psychométriques que nous avons choisi de retenir pour notre étude de l'adéquation des tests spatiaux pour des collégiens de sixième, soient le SBST, le MRT et le CFT. Ayant observé des différences de comportement chez les élèves parisiens entre la prise de données en début d'année scolaire et celle en milieu d'année scolaire, nous choisissons de retenir les scores recueillis à Paris en mars pour le SBST et le MRT et en mai dans le Nord pour le CFT, pour utiliser des échantillons dont les caractéristiques sont proches.

Nous nous appuyons sur la Figure 1 comme arbre de décision pour choisir les méthodes de validation des propriétés psychométriques les plus appropriées aux caractéristiques des trois tests que nous avons retenus.

3.4.1 Standardisation

Administration : comme décrit dans la méthodologie de la pré-expérimentation et de l'expérimentation, nous avons traduit le SBST et le CFT. Nous avons utilisé la version d'Albaret et Aubert (1998) du MRT. Les temps de réponse ont été respectés à l'exception de celui du SBST qui a été réduit à 15 minutes.

Cotation : la cotation définie par les auteurs a été respectée.

3.4.2 Objectivité

Langage approprié : nous n'avons pas relevé de difficulté concernant les instructions lors de la passation des trois tests.

Temps de réponse approprié : nous ne retenons pas ce critère d'objectivité pour ces tests de vitesse : de précédentes études ont établi que la plus part des répondants adultes ne terminent pas le MRT dans le temps imparti (Hegarty, 2018). S'agissant du CFT, les instructions précisent qu'il n'est pas attendu de répondre aux 49 questions dans les 10 minutes imparties. Concernant le SBST, nous constatons que 99 % des élèves parviennent à le terminer lors de la mesure de mars.

3.4.3 Validité

Validité de contenu : par manque de temps, nous n'avons pas inclus d'investigation des procédés cognitifs mobilisés par nos répondants dans ces expérimentations, dont l'objectif premier était d'étudier l'impact d'enseignements technologiques sur les scores spatiaux d'élèves de sixième (Charles et Jaillet, 2023). L'échantillon restreint de quatre élèves âgés de 12 ans et consultés pour tester le protocole expérimental avant la pré-expérimentation, avait été interrogé dans des entretiens rétrospectifs pour investiguer les processus de résolution. Les répondants y avaient décrit des stratégies correspondant aux habiletés visées par le test (e.g. transformation mentale pour le SBST), des stratégies différentes des habiletés visées par le test (e.g. utilisation de la symétrie dans une question du SRI visant le changement de point de vue) et des combinaisons de stratégies (e.g. changement de point de vue et transformation mentale pour le SBST).

Validité de critère : de précédentes études (Shea *et al.*, 2001 ; Wai *et al.*, 2009) ayant établi la prédictivité de la performance spatiale mesurée à l'adolescence de la performance académique et professionnelle en STIM, nous optons pour l'étude de corrélations entre les scores spatiaux et les moyennes de mathématiques, et entre les scores spatiaux et les moyennes de sciences et technologie.

Nous optons pour des corrélations par rangs de Spearman, adaptées aux distributions ne respectant pas la loi normale (Gaudron, 2016). Les corrélations observées, détaillées dans le Tableau 5, sont très significatives ($p < 0{,}01$) pour les trois tests.

Tableau 5 : Statistiques descriptives et corrélations de Spearman pour les scores spatiaux et les moyennes annuelles de mathématiques et de sciences et technologie[6]

Variable	*N*	*M*	*ET*	Min	Max	1	2	3	4	5	6	7
1. MRT – P	92	8,20	4,62	0	20	—	—	—	—	0,33**	—	0,48***
2. SBST – P	92	15,35	5,35	3	27	—	—	—	—	0,32**	—	0,38***
3. CFT – N	80	30,81	11,28	8	63	—	—	—	0,55***	—	0,54***	—
4. Math – N	90	15,30	3,03	7,65	19,56	—	—	0,55***	—	—	—	—
5. Math – P	92	13,55	4,33	1	19,75	0,33**	0,32**	—	—	—	—	—
6. ST - N	80	14,96	2,35	8,20	18,80	—	—	0,54**	—	—	—	—
7. ST – P	92	13,39	3,50	5,26	18,41	0,48***	0,38***	—	—	—	—	—

Nous étudions les nuages de points, illustrés dans les Figures 6A-F, pour vérifier les corrélations observées. Ceux-ci confirment les coefficients de corrélation et montrent des dispersions de points plus resserrées autour de la droite de corrélation lorsqu'il s'agit des moyennes de sciences et technologie (Figures 6 B, D et F). Nous remarquons cependant pour les nuages de points relatifs à la relation avec les mathématiques (Figures 6 A, C et E), que, malgré une tendance générale, un nombre d'individus, variable en proportion selon la corrélation, s'écarte de cette tendance, ce qui est cohérent avec des coefficients de corrélation significatifs mais inférieurs à 0,50 (Hopkins, 1998).

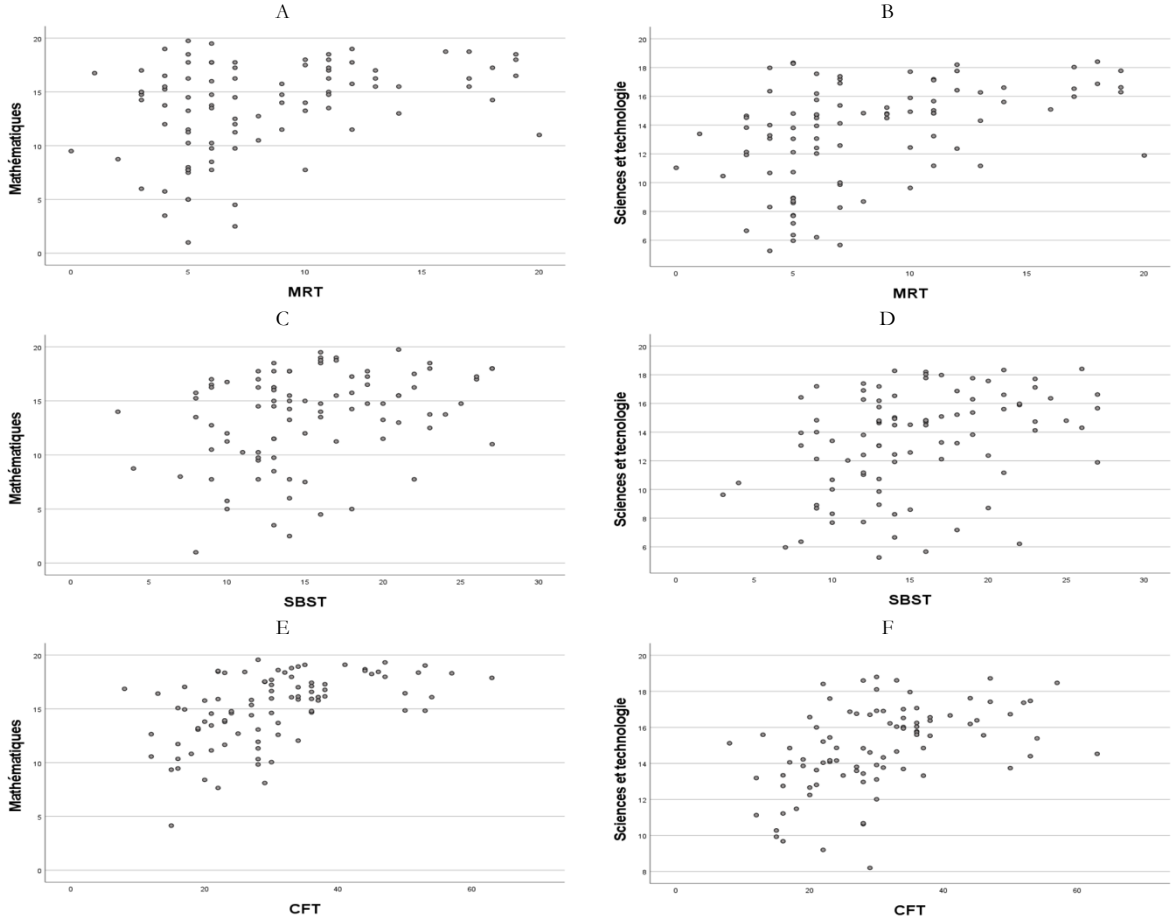

Figure 6 : Nuages de points de la dispersion du MRT, du SBST et du CFT en fonction du niveau de mathématiques (A, C et E) et de sciences et technologie (B, D et F)[7].

[6] *N* = nombre de sujets ; M = moyenne ; ET = écart type ; Maths = mathématiques ; ST = Sciences et technologie ; N = Nord ; P = Paris. ** p<0,01 ; ***p<0,001.

[7] Les Figures A-D concernent les données relevées à Paris, et les Figures E-F celles relevées dans le Nord.

Validité de construit : nous cherchons à vérifier l'adéquation du MRT, SBST et CFT, qui ont été conçus pour des adultes, avec de jeunes adolescents par manque de tests spatiaux conçus pour cette population. Nous ne pouvons donc pas les comparer avec des tests équivalents. Nous ne pouvons les comparer entre eux non plus car ils visent à mesurer des habiletés différentes.

Validité d'apparence : nous n'avons pas relevé auprès de nos sujets de commentaires spontanés remettant en cause la validité du test à leurs yeux.

3.4.4 Fidélité

Nous adaptons la méthodologie de vérification de la fidélité en fonction des caractéristiques des tests. Nous écartons la méthode de la cohérence interne pour le MRT et le CFT qui sont des tests de vitesse (Hopkins, 1998).

Concernant le MRT, nous utilisons les cinq premières réponses de la partie 1 et de la partie 2 pour éviter les items placés en fin de test qui risquent de ne pas avoir été traités par manque de temps. Nous privilégions la méthode de la bissection pour éviter un éventuel effet d'entraînement (American Psychological Association, s.d.d). Nous l'adaptons en créant deux formes équivalentes ($items_1$ 1, 11, 4, 14, 6, 16 ; $items_2$ 2, 12, 3, 13, 5, 15), composées pour la moitié de questions de la partie 1 et pour l'autre de questions de la partie 2, car les distracteurs des items de la partie 2 sont différents de ceux de la première partie (Vandenberg et Kuse, 1978, p. 599). Nous obtenons des moyennes et des écarts types (M_1 = 3,21 ; ET_1 = 1,49 ; M_2 = 2,90 ; ET_2 = 1,52) proches pour les deux formes équivalentes créées (Tableau 6). Le coefficient d'équivalence, calculé au moyen d'une corrélation de rangs de Spearman, est de 0,78 ($p < 0{,}001$). Le nuage de points, illustré dans la Figure 7A, décrit une corrélation des deux moitiés constituées resserrée autour de la droite de corrélation qui confirme la corrélation.

Tableau 6 : Statistiques descriptives et corrélation de rangs de Spearman pour les scores des formes équivalentes du MRT et pour les deux prises de mesure du CFT[8]

Variable	*N*	*M*	*ET*	1	2	3	4
1. MRT Forme 1 - Paris	92	3,21	1,49	—	0,78***	—	—
2. MRT Forme 2 - Paris	92	2,90	1,52	0,78***	—	—	—
3. CFT Recueil 1 - Nord	80	30,81	11,28	—	—	—	0,77***
4. CFT Recueil 2 - Nord	80	39,14	16,45	—	—	0,77***	—

[8] *N* = nombre de sujets ; M = moyenne ; ET = écart type ; * $p<0{,}05$; ** $p<0{,}01$; *** $p<0{,}001$.

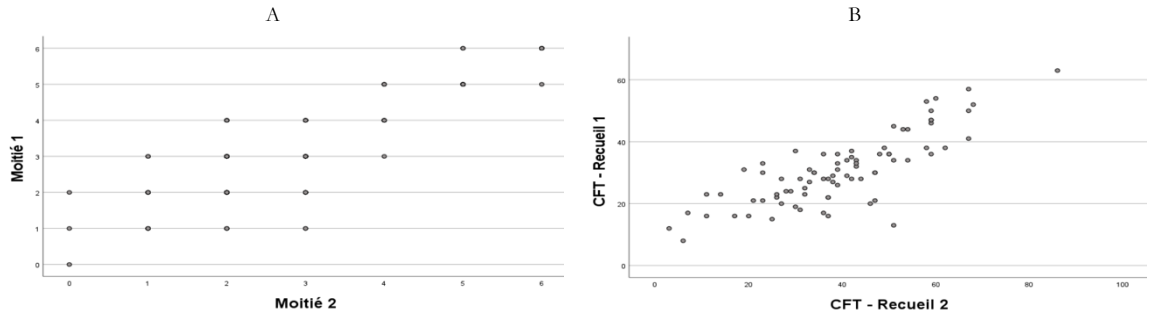

Figure 7 : Nuages de points de la dispersion des scores des moitiés reconstruites du MRT, mesuré à Paris, (A) et des deux mesures du CFT, mesuré dans le Nord, (B)

Concernant le SBST, nous optons pour la méthode de la cohérence interne. Nous obtenons un alpha de Cronbach (1951) de 0,81 (N = 92).

Concernant le CFT, nous optons pour la méthode de la stabilité dans le temps en calculant la corrélation entre les deux prises de mesure du CFT que nous avions réalisées lors de la pré-expérimentation dans le collège nordiste. Le coefficient de stabilité, calculé au moyen d'une corrélation de rangs de Spearman, est de 0,77 (p < 0,001) (Tableau 6). Le nuage de points, illustré dans la Figure 7B, décrit une corrélation resserrée des deux mesures autour de la droite de corrélation qui confirme la corrélation.

3.4.5 Sensibilité

Nous analysons les courbes de distribution des tests spatiaux (Figures 8A-8C), dont un respect de la loi normale indique une difficulté adéquate avec l'échantillon utilisé et une répartition des scores de tous les niveaux de performance suffisante (Hopkins, 1998). Nous complétons cette analyse par celle des moyennes et de l'écart type (Bernaud, 2014), décrits dans le Tableau 5.

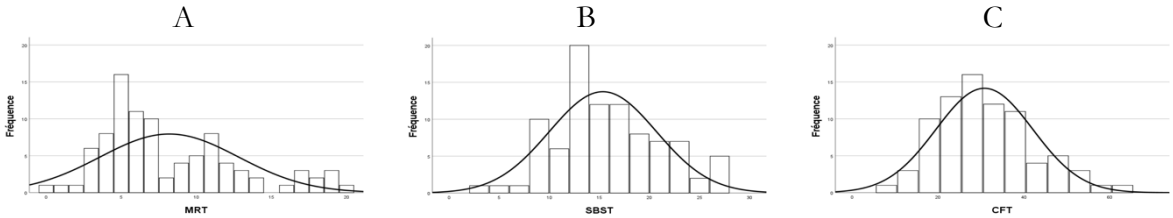

Figure 8 : Courbes de distribution du MRT (Figure 8A), du SBST (Figure 8B) et du CFT (C). A et B concernent les données relevées à Paris et C celles relevées dans le Nord

La distribution des scores du MRT, illustrée dans la Figure 8A, est trimodale et l'écart type de 4,62 est élevé pour 20 questions.

La distribution des scores du SBST, illustrée dans la Figure 8B, est gaussienne, avec une concentration des performances autour des scores compris entre 12 et 18. La moyenne se situe à 15,35, et on peut expliquer un écart-type de 5,35 par le fait que le test comporte 30 questions.

La distribution des scores du CFT, illustrée dans la Figure 8C, est gaussienne, avec une concentration des performances autour des scores compris entre 15 et 39.

4. Discussion

Les quatre tests sélectionnés pour nos pré-expérimentation et expérimentation ont été soumis au procédé de validation des items de Bernaud (2014) (Figure 2). De plus, les propriétés psychométriques du SBST, du MRT et du CFT ont été analysées selon les procédés relevés dans la littérature (Figure 1). Nous discutons tout d'abord les résultats communs obtenus pour le SBST, le MRT et le CFT, avant d'aborder les spécificités des quatre tests expérimentés.

S'agissant du SBST, du MRT et du CFT :

- Nous n'avons pas relevé de difficulté concernant les instructions lors de la passation et en concluons qu'elles sont appropriées pour nos sujets ;
- Selon Bernaud (2014), les indices de difficulté relevés sont acceptables (DSBST = 0,43 ; DMRT = 0,20 ; DCFT = 0,16) ;
- Bien que nous n'ayons pas interrogé les élèves sur leurs stratégies de résolution, nos précédents travaux (Charles, 2023, p. 201-205) confirment le recours à des stratégies alternatives décrites dans la littérature pour le MRT par des lycéens (Albaret et Aubert, 1996) et des adultes (Hegarty, 2018). Ceci est confirmé avec l'échantillon restreint que nous avons consulté avant la pré-expérimentation. Plutôt que mesurer une capacité spécifique, les tests spatiaux expérimentés mobilisent différentes stratégies permettant de résoudre des problèmes spatiaux (Hegarty, 2018)
- Nous n'avons pas pu retenir le critère temps de réponse pour le CFT et le MRT, qui sont des tests de vitesse. Ceci pose le problème des questions en fin de test restées sans réponse, non pas parce que le sujet ne sait pas y répondre mais parce qu'il n'a pas eu le temps d'y répondre (Hopkins, 1998), ainsi que le risque que le répondant priorise une des deux injonctions, c'est-à-dire répondre correctement ou dans le temps imparti, ce qui pourrait produire un score qui ne relève pas de la compétence mesurée. D'autre part, le fait de ne pas répondre à une partie des questions limite les procédés de vérification de la fidélité (Hopkins, 1998) : la méthode de la stabilité dans le temps soulève le problème de l'effet d'entrainement (Scharfen *et al.*, 2018) et la méthode de la bissection nécessite des formes parallèles, qui n'existent pas à notre connaissance pour les tests spatiaux. Nous avons présenté dans cet article l'alternative d'adapter la méthode de la bissection pour le MRT car le test est constitué de deux parties, séparées par une pause, dont nous avons extrait des réponses de manière à utiliser des items de même format (Vandenberg et Kuse, 1978) et en écartant les questions passées en fin de test.
- Les coefficients de corrélation avec les moyennes de mathématiques et de sciences et technologie (Tableau 5) sont proches de la valeur prédictive de 0,50 décrite par Bernaud (2014) et confirmées par l'analyse des nuages de points (Figures 6A-F). Elles sont conformes aux études sur la prédictivité de la performance spatiale mesurée à l'adolescence de la performance scolaire en STIM (Wai *et al.*, 2009). Nous proposons d'en conclure que la validité critérielle est vérifiée pour les trois tests.
- L'alpha de Cronbach de 0,81 ($N = 92$) obtenu pour le SBST, le coefficient d'équivalence des deux moitiés du MRT de 0,78 ($p < 0{,}001$) ($N = 92$) et le

coefficient de stabilité dans le temps du CFT de 0,77 ($p < 0{,}001$) ($N = 80$) correspondent à une fidélité correcte (Bernaud, 2014).

- La distribution gaussienne des scores du SBST et du CFT indique une difficulté adéquate avec l'échantillon utilisé et une répartition des scores de tous les niveaux de performance suffisante (Hopkins, 1998).

Nous abordons ci-après les résultats spécifiques aux quatre tests expérimentés.

4.1. SRI

Le test des consignes a mis en évidence une difficulté de compréhension pour 4 % de notre échantillon. Ce test présente la particularité d'utiliser des questions contextualisées (Ramful *et al.*, 2017) spécifiques à chaque item. La corrélation très significative ($p < 0{,}001$) avec le niveau de français et le coefficient de corrélation ($p = 0{,}58$), supérieur à la valeur prédictive d'un test psychométrique de la performance scolaire habituellement observée (Bernaud, 2014), suggèrent que ce test est adapté pour mesurer les habiletés spatiales d'élèves dont le niveau de français est suffisant pour ne pas limiter la compréhension des instructions.

4.2. SBST

Pour des raisons de disponibilité de nos sujets, nous avons diminué le temps de réponse à 15 minutes. 99 % des élèves parisiens parviennent à le terminer lors de la mesure de mars, ce qui est supérieur aux 90 % de répondants qui doivent avoir terminé l'épreuve dans le temps imparti selon Hopkins (1998). Nous en concluons que notre restriction de temps est adaptée pour des élèves de sixième suffisamment avancés dans leur nouvelle posture, c'est-à-dire capable de se concentrer pendant l'explication d'instructions et de maintenir leur attention pendant une tâche, quand le temps disponible pour les expérimentations est restreint.

4.3. MRT

La courbe de distribution trimodale observée (Figure 8A) semblerait indiquer que notre échantillon soit composé de trois groupes de performance distincts. Nous observons dans le détail des réponses, un groupe de 20 élèves (22 %) qui ne répond pas aux cinq dernières questions de la première partie du test. Un examen approfondi de ce groupe indique qu'il est constitué pour 70 % ($N = 14$) de filles, ce qui correspond à de précédentes études mettant en évidence une performance au MRT inférieure pour les filles, notamment en raison de la priorité qu'elles accordent à la précision de leurs réponses, plutôt qu'à l'injonction de répondre dans le temps imparti (Voyer *et al.*, 2004). Nous comparons la performance aux deux moitiés reconstruites en regard du genre : le test de Mann-Whitney (Mann et Whitney, 1947) produit un résultat non significatif pour la première moitié ($p = 0{,}12$) et significatif ($p < 0{,}5$) pour la seconde. Les boîtes à moustaches, illustrées en Figure 9, décrivent une différence en faveur des garçons, dont la distribution est cependant plus étalée que celle des filles. De plus, malgré un resserrement des scores, on remarque des valeurs extrêmes chez les filles, absentes chez les garçons. Nous étudions la relation entre performance en mathématiques et en sciences et technologie et genre pour repérer d'éventuelles variables explicatives. Le test de Mann-Whitney produit un résultat non significatif pour les mathématiques ($p = 0{,}72$) et pour les sciences et technologie ($p = 0{,}70$).

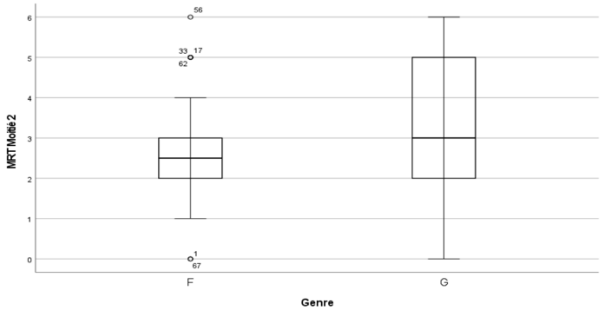

Figure 9 : Distribution des scores de MRT – Moitié 2 selon le genre[9]

4.4. CFT

On remarque que l'indice difficulté du CFT est proche de la limite basse, ce qui peut s'expliquer par le fait que ce test est très limité dans le temps et que le score est impacté par le nombre d'items sans réponses. Le test comporte 49 questions et jusqu'à quatre réponses correctes sont possibles pour certaines questions, ce qui peut expliquer la moyenne de 30,81 et l'écart-type de 11,28 points : il se peut que la stratégie de réponse des élèves (e.g. beaucoup d'items traités avec une réponse unique ou au contraire peu d'items répondus mais avec plusieurs réponses), ait un effet sur le score total.

5. Limites et perspectives

Bernaud (2014) recommande d'utiliser des échantillons conséquents, i.e. compris entre 300 et 500 sujets, et représentatifs pour valider les items d'un test psychométrique. Les échantillons dont nous avons exploité les données sont inférieurs à la centaine de répondants et leurs caractéristiques sont proches, néanmoins différentes : de manière à conserver une certaine homogénéité d'âge et de maturité scolaire, nous avons utilisé les données de la deuxième prise de données en mars de l'échantillon parisien, alors que nous avons pris les scores collectés au premier recueil dans le Nord en mai. Il se peut que le fait d'avoir passé le test deux fois ait un effet sur la performance des élèves parisiens : dans leur méta-analyse, Scharfen *et al.* (2018) ont mis en évidence des effets d'entraînement[10] significatifs entre la première et la seconde administration de tests cognitifs, avec un plus grand effet pour les tests contenant des tâches utilisant des contenus figuraux, c'est-à-dire des images, des objets, des matériaux abstraits ou spatiaux, que pour les tests numériques. Nous ne pouvons cependant pas contrôler ce résultat en comparant la performance de ces deux groupes, car les temps de réponse alloués sont différents. D'autre part, un test de Mann-Whitney indique des différences de performance significatives pour les mathématiques ($p = 0{,}001$) et pour les sciences et technologie ($p = 0{,}005$) entre les deux collèges : il est cependant difficile d'interpréter ces résultats étant donné que les évaluations sont différentes, d'une part, et que les sciences et technologie sont enseignées par le même enseignant dans le collège nordiste et par trois enseignants différents dans le collège parisien, d'autre part. Bernaud (2014, p. 80) indique cependant « *qu'il est fondamental de*

[9] F = Filles, G = Garçons. La distribution des scores des filles de la seconde moitié reconstruite du MRT est caractérisée par une performance relativement homogène et quelques valeurs extrêmes très élevées et très faibles. Celle des garçons montre un étalement important des scores, soit une performance hétérogène.

[10] « *any change or improvement that results from practice or repetition of task items or activities* [tout changement ou amélioration qui résulte de la pratique ou de la répétion de tâches ou d'activités] » (American Psychological Association, s.d.d)

préserver une certaine parenté (du point de vue démographique) entre l'échantillon concerné par la phase de validation des items et celui de la phase de validation des échelles ». Notre limitation du temps du SBST à 15 minutes, pour des raisons de disponibilité des échantillons, pourrait de plus remettre en cause la validité de construit de ce test : les écarts de performance mesurés entre les individus pourraient être dus, non pas à leur différence de performance au construit mesuré, mais aux erreurs de mesure qui éloignent la performance mesurée des scores vrais (Maeda et Yoon, 2013). Notre distribution suivant une distribution gaussienne, nous suggérons cependant que la limite de 15 minutes permet de mesurer une différence de performance acceptable, quand le temps disponible pour les expérimentations est restreint. De nouvelles expérimentations sont nécessaires pour évaluer l'impact de cette modification sur la performance des élèves. En l'état, elle ne nous permet pas de comparer nos résultats à des études respectant le temps défini par les auteurs.

6. Conclusion

Évaluer les habiletés spatiales de jeunes adolescents nécessite d'identifier des outils de mesure appropriés à ces sujets. Notre revue de littérature nous a permis de construire un référentiel d'analyse des tests spatiaux, d'une part, et d'évaluation des qualités d'un test psychométrique, d'autre part. Ceci nous a permis de sélectionner et d'expérimenter quatre tests spatiaux, i.e. le SRI, le MRT, le SBST et le CFT, auprès de deux échantillons correspondant à notre public cible. Notre étude avait pour ambition de tester la validité du SRI et du SBST dans un temps de passation réduit et celles du MRT et du CFT auprès d'élèves de sixième. Nos résultats décrivent une corrélation très significative ($p < 0{,}001$) entre scores de SRI et moyenne annuelle de français, qui suggère que ce test mesure plus d'un construit. Les différentes analyses que nous avons menées pour vérifier les qualités métrologiques des tests sélectionnés, ainsi que l'étude de la validité de leurs items pour nos échantillons, nous invitent à conclure à l'adéquation du SBST, avec un temps de passation réduit à 15 minutes, pour des sixièmes en fin d'année scolaire et du CFT pour notre public cible, quand le temps disponible pour les expérimentations est restreint. Bien que la distribution trimodale du MRT interroge sur les causes de différences de performance observées, dont on sait que le genre est un facteur chez les adolescents (Albaret et Aubert, 1996 ; Hoyek *et al.*, 2012 ; Jansen-Osmann et Heil, 2007), nous n'écartons pas l'intérêt de ce test pour mesurer les habiletés spatiales de jeunes adolescents : ce test permet justement d'identifier des pistes d'explication de différences de performance, ici le genre, et des âges auxquels ces différences émergent (Hoyek *et al.*, 2012). Les techniques d'analyse des tests spatiaux et de leurs qualités métrologiques que nous avons conçues nous ont permis une étude heuristique de l'adéquation de tests psychométriques pour un public spécifique. Des recherches complémentaires sont nécessaires pour en établir la robustesse.

7. Références bibliographiques

Agbanglanon, S. (2019). *Outils numériques dans l'apprentissage de la conception mécanique : analyse des liens entre représentations externes et capacités visuo-spatiales dans le processus de conception* [thèse de doctorat, Université de Cergy Pontoise, Cergy-Pontoise, France]. https://tel.archives-ouvertes.fr/tel-02623908

Albaret, J. M. et Aubert, E. (1996). Étalonnage 15-19 ans du test de rotation mentale de Vandenberg. *EVOLUTIONS psychomotrices, 8*(34), 269-278.

American Psychological Association. (s.d.-a). True score. Dans *APA Dictionary of Psychology*. https://dictionary.apa.org/true-score

American Psychological Association. (s.d.b). Performance test. *Dans APA Dictionary of Psychology*. https://dictionary.apa.org/performance-test

American Psychological Association. (s.d.c). Power test. Dans *APA Dictionary of Psychology*. https://dictionary.apa.org/power-test

American Psychological Association. (s.d.d). Practice effect. Dans *APA Dictionary of Psychology*. https://dictionary.apa.org/practice-effect

American Psychological Association. (s.d.e). Standardized test. Dans *APA Dictionary of Psychology*. https://dictionary.apa.org/standardized-test

Bernaud, J.-L. (2014). *Méthodes de tests et questionnaires en psychologie*. Dunod.

Branoff, T. (2000). Spatial Visualization Measurement: A Modification of the Purdue Spatial Visualization Test - Visualization of Rotations. *Engineering Design Graphics Journal, 64*(2), 14-22.

Branoff, T. et Dobelis, M. (2012). The Relationship between Spatial Visualization Ability and Students' Ability to Model 3D Objects from Engineering Assembly Drawings. *Engineering Design Graphics Journal, 76*(3), 37-43.

Brown, W. (1910). Some Experimental Results in the Correlation of Mental Abilities1. *British Journal of Psychology, 1904-1920, 3*(3), 296-322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Charles, S. (2023). *Habileté spatiale et stratégies de modélisation 3D* [thèse de doctorat, CY Cergy Paris Université, Cergy-Pontoise, France]. https://hal.science/tel-04097396

Charles, S. et Jaillet, A. (2023, 3 avril). *Apprend-on en passant des tests ?* [communication orale]. 34ème colloque de l'ADMEE-Europe, Université de Mons, Mons, Belgique.

Charles, S., Jaillet, A., Peyret, N. et Jeannin, L. (2019). Éléments de mesure de la compétence de visualisation spatiale d'étudiants ingénieurs en mécanique. Dans *Actes du CFM 2019*. https://cfm2019.sciencesconf.org/244039

Cohen, C. A. et Hegarty, M. (2007). Sources of Difficulty in Imagining Cross Sections of 3D Objects. Dans *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (p. 179-184).

Cohen, C. A. et Hegarty, M. (2012). Inferring cross sections of 3D objects: A new spatial thinking test. *Learning and Individual Differences, 22*, 868-874.

Cortina, power. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1949). *Essentials of psychological testing* (p. xiii, 475). Harper.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

Eliot, J. (1983a). Historical background. Dans *An international directory of spatial tests* (p. 1-10). NFER-Nelson.

Eliot, J. (1983b). The classification of spatial tests. Dans *An international directory of spatial tests* (p. 11-15). NFER-Nelson.

Eliot, J. et Macfarlane Smith, I. (1983). *An international directory of spatial tests*. NFER-Nelson.

Ferguson, G. A. (1949). On the theory of test discrimination. *Psychometrika, 14*(1), 61-68. https://doi.org/10.1007/BF02290141

Ferrando Piera, P. J. (2012). Assessing the discriminating power of item and test scores in the linear factor-analysis model. *Psicológica: Revista de metodología y psicología experimental*, 33(1), 65-85.

Gaudron, J. P. (2016). *R Commander : Petit guide pratique 1. Statistiques de base.*

Hegarty, M. (2018). Ability and sex differences in spatial thinking: What does the mental rotation test really measure? *Psychonomic Bulletin & Review, 25*(3), 1212-1219. https://doi.org/10.3758/s13423-017-1347-z

Hopkins, K. D. (1998). *Educational and Psychological Measurement and Evaluation* (8e éd.). Allyn & Bacon.

Hoyek, N., Collet, C. et Guillot, A. (2010). Représentation mentale et processus moteur : le cas de la rotation mentale. *Science & Motricité,* (71), 29-39. https://doi.org/10.1051/sm/2009013

Hoyek, Nady, Collet, C., Fargier, P. et Guillot, A. (2012). The Use of the Vandenberg and Kuse Mental Rotation Test in Children. *Journal of Individual Differences, 33*(1), 62-67. https://doi.org/10.1027/1614-0001/a000063

ISAE-Supméca. (2016). *Expérimenter l'Apprentissage par problèmes et Projets via la conception 3D / EXAPP_3D* [note de synthèse].

Jansen-Osmann, P. et Heil, M. (2007). Developmental aspects of parietal hemispheric asymmetry during mental rotation. *NeuroReport, 18*(2), 175-178. https://doi.org/10.1097/WNR.0b013e328010ff6b

Kuder, G. F. et Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160. https://doi.org/10.1007/BF02288391

Le Corff, Y., Yergeau, E., Beaudin, M.-E. et Dorceus, S. (2017). Psychométrie. *Site Psychométrie à l'UdeS*. http://psychometrie.espaceweb.usherbrooke.ca/instrument-psychometrique

Lohman, D. F., Pellegrino, J. W., Alderton, D. L. et Regian, J. W. (1987). Dimensions and Components of Individual Differences in Spatial Abilities. Dans S. H. Irvine et S. E. Newstead (dir.), *Intelligence and Cognition: Contemporary Frames of Reference* (p. 253-312). Springer Netherlands. https://doi.org/10.1007/978-94-010-9437-5_6

Maeda, Y. et Yoon, S. Y. (2013). A Meta-Analysis on Gender Differences in Mental Rotation Ability Measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R). *Educational Psychology Review, 25*(1), 69-94. https://doi.org/10.1007/s10648-012-9215-x

Mann, H. B. et Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics, 18*(1), 50-60. https://doi.org/10.1214/aoms/1177730491

Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche. (2015). *Bulletin officiel spécial n° 11 du 26 novembre 2015*.

Ramful, A., Lowrie, T. et Logan, T. (2017). Measurement of Spatial Ability: Construction and Validation of the Spatial Reasoning Instrument for Middle School Students. *Journal of Psychoeducational Assessment, 35*(7), 709-727. https://doi.org/10.1177/0734282916659207

Scharfen, J., Peters, J. M. et Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence, 67*, 44-66. https://doi.org/10.1016/j.intell.2018.01.003

Shapiro, S. S. et Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika, 52*(3/4), 591-611. https://doi.org/10.2307/2333709

Shea, D. L., Lubinski, D. et Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year Longitudinal Study. *Journal of Educational Psychology, 93*(3), 604-614. https://doi.org/10.1037/0022-0663.93.3.604

Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3(3), 271-295. https://doi-org.bibdocs.u-cergy.fr/10.1111/j.2044-8295.1910.tb00206.x

Steinhauer, H. M. (2012). Correlation Between a Student's Performance on the Mental Cutting Test and Their 3D Parametric Modeling Ability. *Engineering Design Graphics Journal, 76*(3), 44-48.

Tartre, L. A. (1990). Spatial orientation skill and mathematical problem solving. *Journal for Research in Mathematics Education*, 216-229.

Thurstone, L. L. et Jeffrey, T. E. (1956). *Closure Flexibility (Concealed Figures) Test - Form A*. Industrial Relations Center - The University of Chicago.

Vandenberg, S. G. et Kuse, A. R. (1978). Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization. *Perceptual and Motor Skills, 47*(2), 599-604.

https://doi.org/10.2466/pms.1978.47.2.599

Voyer, D., Rodgers, M. A. et McCormick, P. A. (2004). Timing conditions and the magnitude of gender differences on the Mental Rotations Test. *Memory & Cognition, 32*(1), 72-82. https://doi.org/10.3758/BF03195821

Wai, J., Lubinski, D. et Benbow, C. P. (2009). Spatial Ability for STEM Domains: Aligning over 50 Years of Cumulative Psychological Knowledge Solidifies Its Importance. *Journal of Educational Psychology, 101*(4), 817-835. https://doi.org/10.1037/a0016127

Yue, J. (2004). Spatial visualization by orthogonal rotations. Dans *Proceedings of ASEE Annual Conference and Exposition* (vol. 9, p. 1-10).

Yue, J. (2006). Spatial Visualization by Isometric Drawing. Dans *Proceedings of the 2006 IJME - INTERTECH Conference* (vol. 3).

8. Annexes

Tableau 7 : Synthèse des caractéristiques des tests spatiaux

Caractéristique	**État**		
Compétence visée	Unique	Multiples	
Temps de réponse	Très limité	Peu limité	Libre
Instructions	Communes à tous les items	Spécifiques à chaque item	
	Présence de question(s) d'entraînement		
	Normalisées		
	Contextualisées		
Réponse	Choix parmi des alternatives	Libre	Dessin
Stimuli	Abstrait	Familier	
	Objet manipulable	Photographie	Dessin
	Dessin axonométrique	Dessin isométrique	Dessin libre
	Couleur	Noir et blanc	
	2D	3D	
Âge visé ou investigué	Enfants	Adolescents	Adultes
Format	Manipulation de matériel	Papier-crayon	Informatisé
Difficulté des items	Croissante	Équivalente	Variable
Disponibilité	Commercialisé	Disponible	Épuisé

Tableau 8 : Synthèse des indicateurs de qualité des tests spatiaux relevés dans la littérature de validation de tests spatiaux et de leurs outils d'évaluation

	Étude			
	Albaret et Aubert (1996)	**Hoyek *et al.* (2012)**	**Ramful *et al.* (2016)**	**Cohen et Hegarty (2007 ; 2012)**
Objectif	Adéquation du MRT pour lycéens	Adéquation du MRT pour collégiens	Validation du SRI	Validation du SBST
Standardisation	Passation Cotation	Passation Cotation	Passation Cotation	Passation Cotation
Fidélité				
Inter-correcteur	Cotation standardisée	Cotation standardisée	Cotation standardisée	Cotation standardisée
Intra-test	Méthode de bissection			
Erreur-type de mesure	Calcul			
Validité				
de contenu	Théorique Questionnaire posttest	Corrélation avec test 2D *ad hoc*	Théorique Alignement avec contenus scolaires	Théorique
de construit	Méthode de bissection		Corrélation avec tests reconnus et entre les sous-échelles Index de séparation	Corrélation du test avec tests reconnus et entre les sous-échelles
Fiabilité			α de Cronbach du test et des sous-échelles	α de Cronbach du test et des sous-échelles
Sensibilité	Calcul de l'erreur-type de mesure		Mesure de la symétrie et de l'acuité	Analyse de la distribution
Influence de caractéristiques individuelles	Genre Âge Spécialisation	Genre Âge	Âge	Genre