Introspection et métacognition :
Les mécanismes de la connaissance de soi

Stanislas Dehaene
Chaire de Psychologie Cognitive Expérimentale

Cours

Notre capacité d’introspection est-elle illusoire?

Vers une typologie du connu et de l'inconnu

"As we know, there are known knowns; there are things we know we know.

We also know there are known unknowns; that is to say we know there are some things we do not know.

But there are also unknown unknowns -- the ones we don't know we don't know."

Donald Rumsfeld (Department of Defense news briefing, *12 février 2002*)

Ainsi laissait-il entendre qu'il existait des choses tellement secrètes que lui-même ne savait même pas qu'il ne les savait pas.

(en l'occurrence la présence d'armes de destruction massive en Irak).

Il était absolument certain de son ignorance!

Mais, en fait, non!

Il ne savait pas qu'il ne savait pas qu'il ne savait pas.

L'idée que nous disposons d'une excellente introspection est-elle illusoire?

- Les données sont contradictoires:
  - Certaines montrent une cécité totale à ses propres mécanismes décisionnels et une incapacité de prédire ses propres performances dans un futur plus ou moins lointain
  - D'autres démontrent une excellente capacité de rapporter ses processus mentaux à un instant donné

- Exemples classiques de cécité de l'introspection
  - Nisbett et Wilson (1977)
  - Certains syndromes neuropsychologiques tels que l'anosognosie pour l'hémiplégie (Berti et al., *Science* 2005)
  - L'ensemble des traitements non-conscients.

  Paradoxalement, l'imagerie cérébrale (« hétéro-spection ») peut être meilleure que le participant lui-même.

  Exemples : Haynes & Rees (2005): décodage de l'orientation visuelle dans V1; Soon & Haynes (2008): décodage de l'intention ~10 secondes avant l'action)

Un exemple de métacognition incorrecte: le sentiment d'être proche de la solution d'un problème

Metcalfe, J. (1986). Premonitions of insight predict impending error. *JEP:LMC, 12(4), 623-634.*

•Résolution d'énigmes tels que le problème des quatre chaînes de 3 anneaux.
• Toutes les 10 secondes, le sujet note sur une échelle de 0 à 10 son sentiment d'être plus ou moins « chaud » ou « froid »
• Ces résultats sont comparés pour les problèmes que la personne a ou n'a pas résolus correctement.
• Résultats:
- non seulement le « réchauffement » n'est pas une bonne indication que la solution est proche
- mais c'est l'inverse: les sujets donnent des valeurs plus « chaudes » avant une erreur qu'avant une réponse correcte !
- Ces résultats sont répliqués inter-sujets (expérience 1) et intra-sujets (expérience 2), avec des problèmes de toutes sortes.

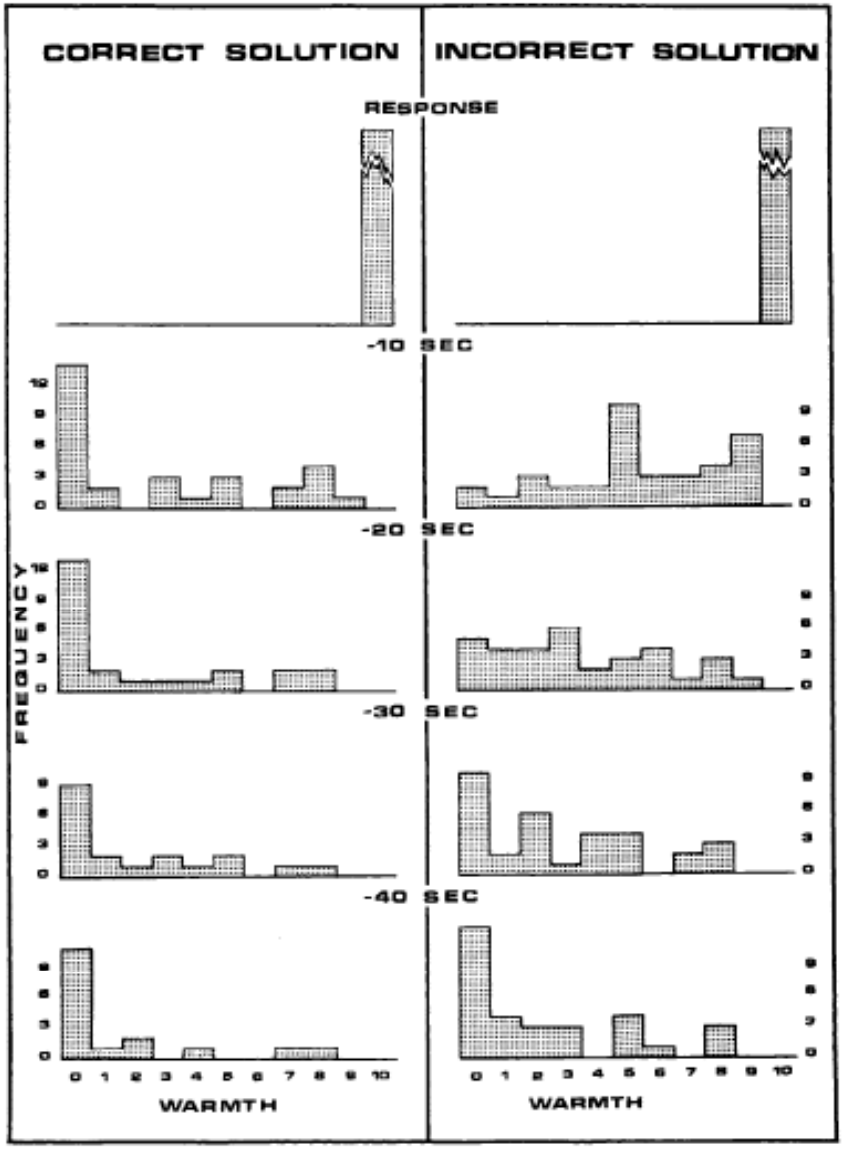
Warmth Ratings on the Last Three Intervals Before Solution for Correct and Incorrect Responses in Experiments 1, 2, 3, 4, and 5

		Interval rating			
Response	*N*	Third last	Second last	Last	Solution
		Experiment 1 (problem)			
Correct	19	2.05	2.42	3.47	10
Incorrect	33	2.92	3.57	5.25	10
		Experiment 2 (problems)			
Correct	29	3.52	3.73	4.60	10
Incorrect	29	4.16	4.64	5.02	10

Un exemple de métacognition incorrecte: le sentiment d'être proche de la solution d'un problème

Metcalfe, J. (1986). Premonitions of insight predict impending error. *JEP:LMC, 12(4), 623-634.*

• l'approche de la solution se caractérise par une découverte abrupte, sans signes avant-coureurs (*insight*).
• Interprétation:
-La découverte est un processus tout-ou-rien et entièrement non-conscient (cf. Poincaré, Hadamard…)
- le sentiment d'approcher traduit en fait l'acceptation progressive d'une solution inélégante.



En général, la métacognition n’est pas totalement fausse

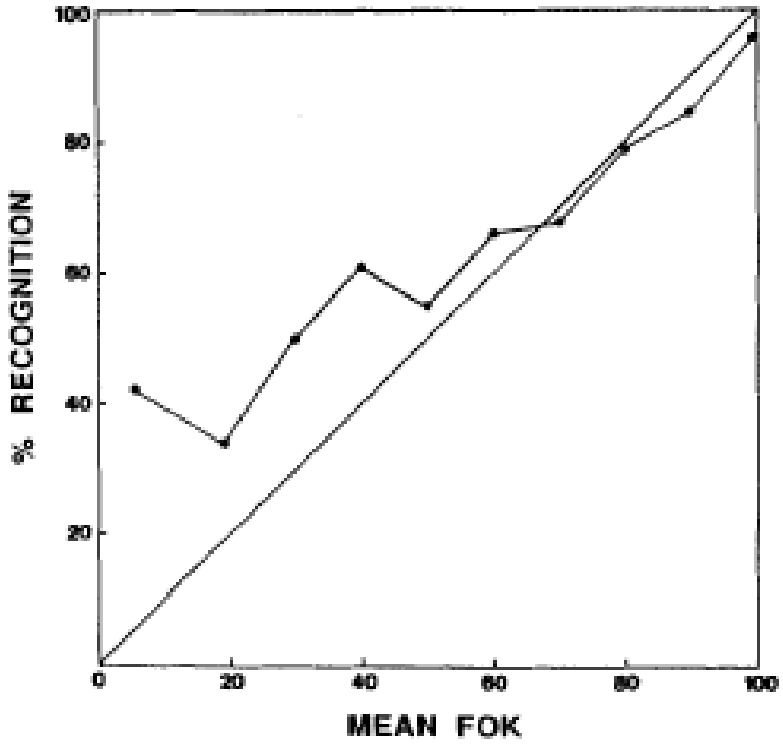
La plupart du temps, les jugements métacognitifs corrèlent faiblement avec les performances de premier ordre.

Par exemple:

- L’expérience du « mot sur le bout de langue » (*Tip of the Tongue* ou *TOT*):
La reconnaissance ultérieure du mot est meilleure lorsque le sujet dit être dans l’état « TOT » (Bennett Schwartz, *Tip of the tongue states*, 2002)

- Le « sentiment de savoir » (*Feeling of knowing*):
La reconnaissance ultérieure d’une chaîne de caractères varie de façon monotone avec le « sentiment de savoir ».

score de reconnaissance
(parmi 8 choix possibles)



Sentiment de savoir (estimation, en pourcentage, de la probabilité de reconnaître la chaîne de caractère apprise)

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychol Rev, 100(4), 609-639.*

Une distinction essentielle: Précision relative versus calibration absolue des jugements

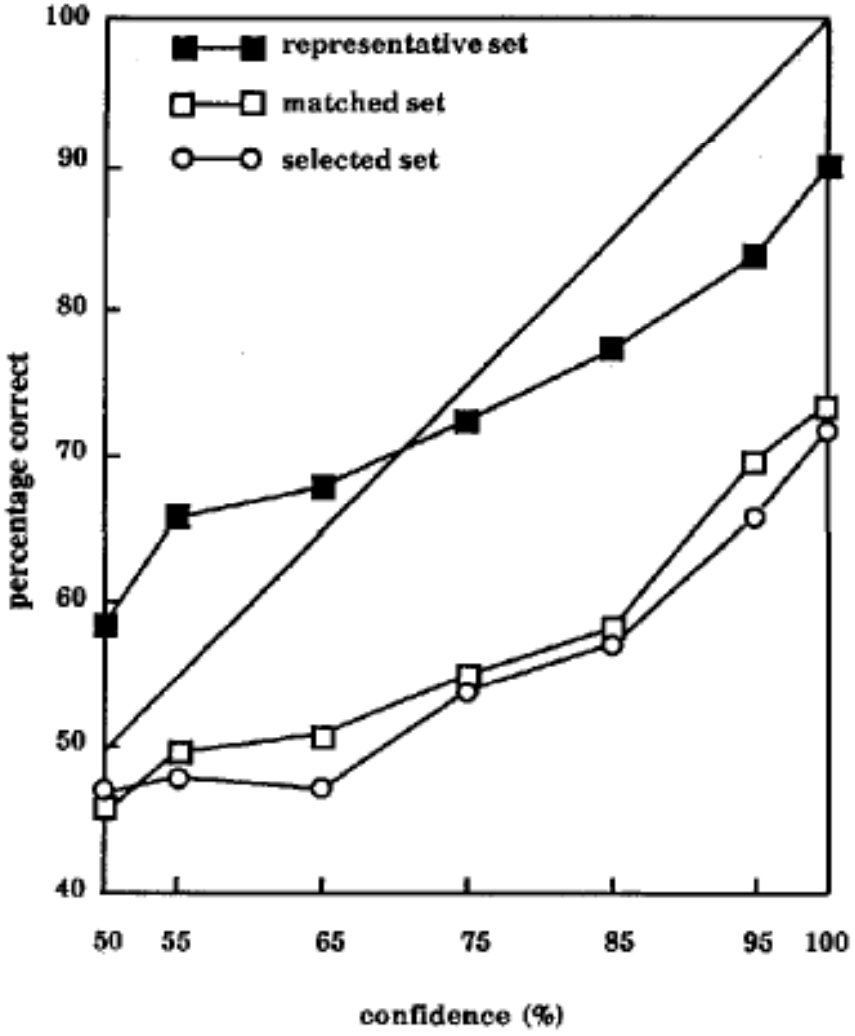
Le jugement métacognitif est généralement corrélé, quoique modestement, avec la performance réelle *(metacognitive accuracy).*

Cependant il n'est pas toujours bien calibré *(metacognitive calibration)*:

- **surestimation** systématique des compétences
- parfois, également, **sous-estimation** de l'intuition pour les items mal maîtrisés.

Ce qui résulte dans l' « effet difficile-facile » (*hard-easy effect)*:

- la mémoire des items faciles est sur-estimée
- la mémoire des items difficiles est sous-estimée



Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychol Rev, 98(4), 506-528.*

Peut-on « calibrer » ses jugements métacognitifs?

- Dans une tâche de méta-mémoire, l'entrainement avec feedback aide, mais là encore, il faut distinguer les deux variables:
- la corrélation entre les jugements et les performances (*JOL accuracy*) augmente
- mais le biais de sous-estimation augmente également!

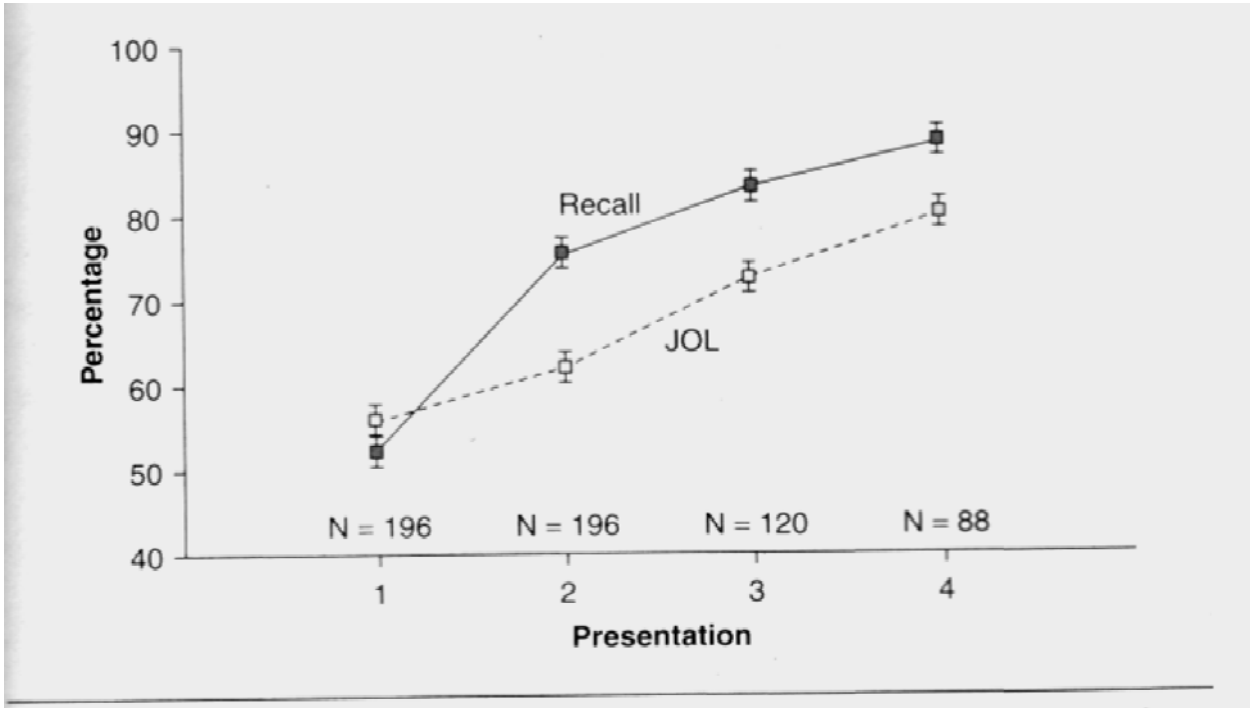


Figure 5.1 The underconfidence-with-practice (UWP) effect. Percentage of correct recall (solid line) and predicted recall (JOL, dotted line) plotted as a function of the number of study presentations. Values were computed by collapsing outcomes from 11 different experiments.

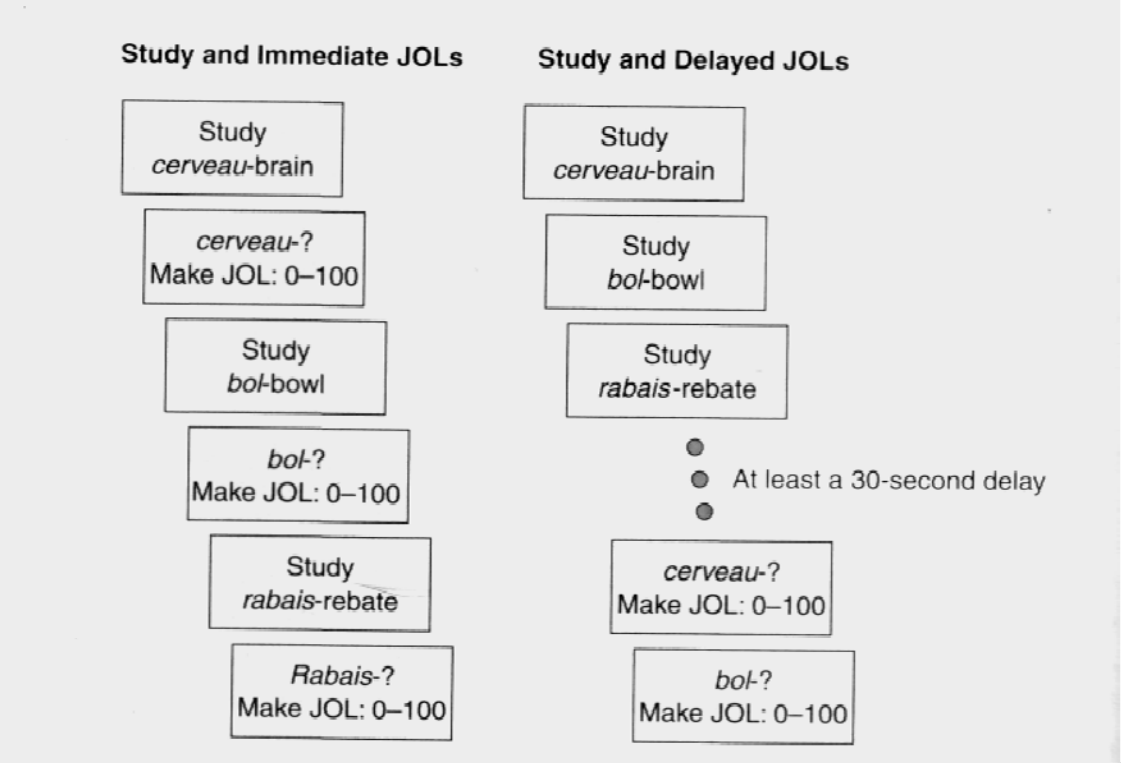
Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *J Exp Psychol Gen, 131(2), 147-162.*

L’introduction d’un délai améliore massivement le *judgement of learning*

- Il est plus facile de juger si l’on a appris correctement après un délai qu’immédiatement après l’apprentissage.

- Corrélation JOL/performance:
- immédiate: +.38
- différée: +.90

Le phénomène est éminemment réplicable.



Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL" effect". *Psychological Science, 2, 267-270.*

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Mem Cognit, 20(4), 374-380.*

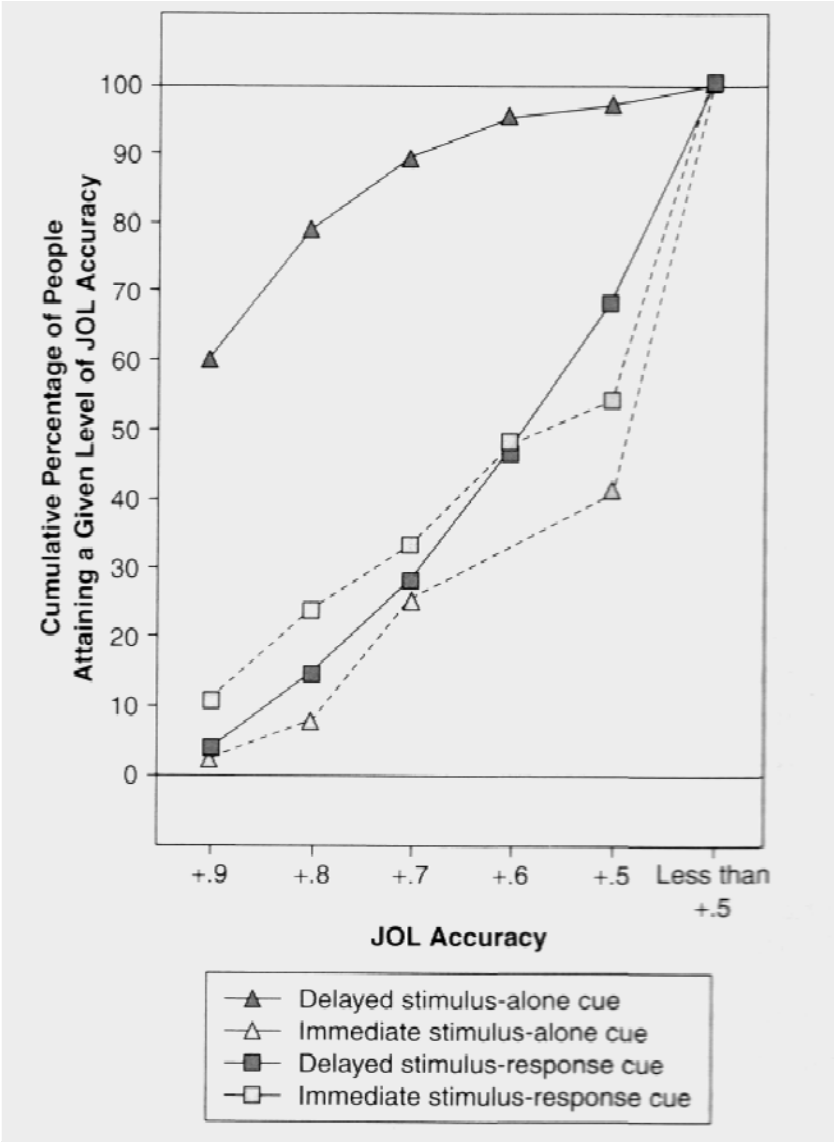
L'introduction d'un délai améliore massivement le *judgement of learning*

- L'explication la plus simple (*Monitoring Dual-Memory*, Nelson &Dunlosky 1991):

Pour répondre au JOL, les sujets essaient de récupérer l'information en mémoire. Ils fondent leur réponse métacognitive sur leur succès ou leur échec – et celui-ci prédit également les performance ultérieures.

A court-délai, la présence de l'information en mémoire à court terme empêche d'évaluer sa propre mémoire.

- Prédiction: l'effet délai doit disparaître si l'on présente à la fois l'indice et la réponse.
- Autre explication possible: *Self-fulfilling prophecy* (Spellman & Bjork 1992): le fait même de tester sa mémoire facilite la mémorisation, .



Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Mem Cognit, 20(4), 374-380.*

Vaut-il mieux étudier ou se tester?

Les expériences de Henry Roediger et collaborateurs

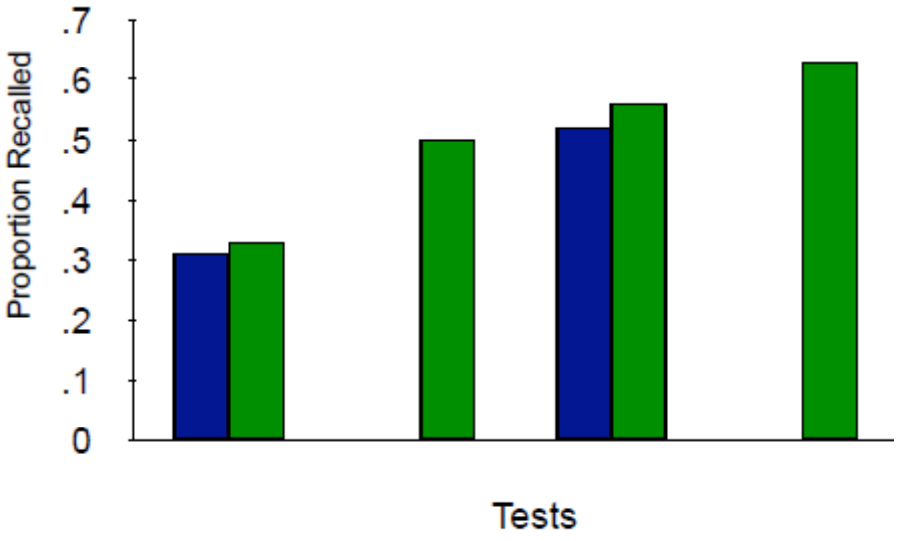
La plupart des études de la consolidation en mémoire consistent en une alternance de moments d'étude et de moments de test: S(tudy) T(est) S T S T S T

Le test est souvent considéré comme une simple mesure, qui ne contribue pas à l'apprentissage.

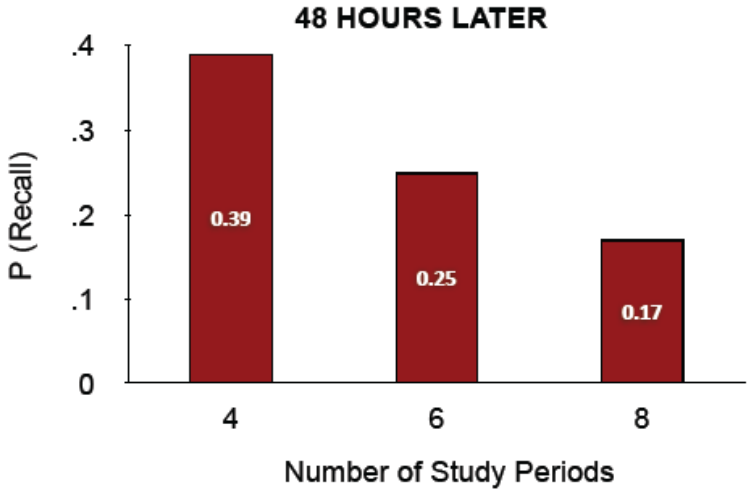
Pour le savoir, Zaromb et Roediger ont manipulé le nombre de périodes d'étude et de test:

ST ST ST ST	4 study, 4 test
ST SS ST SS	6 study, 2 test
SS SS SS SS	8 study, 0 test

Pendant l'apprentissage, résultats similaires…



Mais 48 h plus tard, c'est le nombre de tests qui compte, pas le temps d'étude.



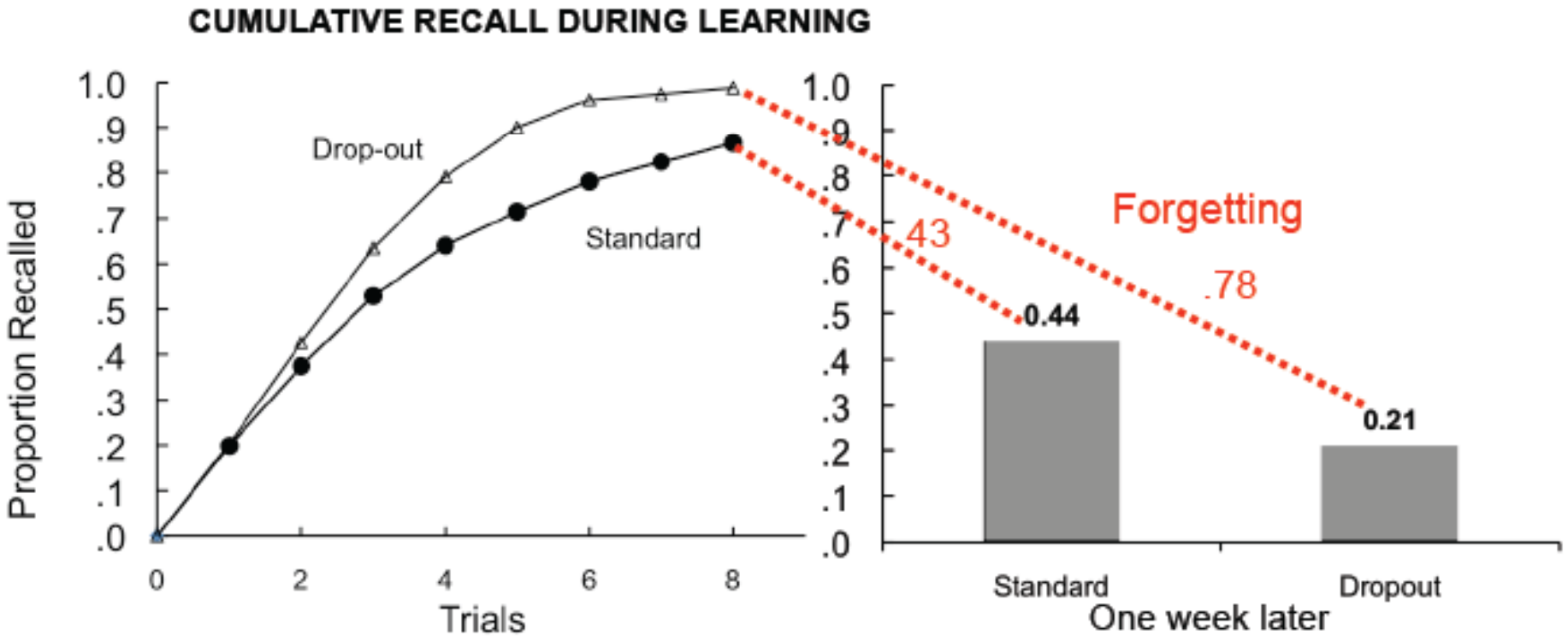
Vaut-il mieux étudier ou se tester?

Les expériences de Henry Roediger et collaborateurs

Pourrait-on améliorer les performances en focalisant l'étude sur les items non-retenus?

Oui durant la période d'apprentissage: apprentissage plus rapide, 100% correct

Non après une semaine: Les deux conditions s'inversent!



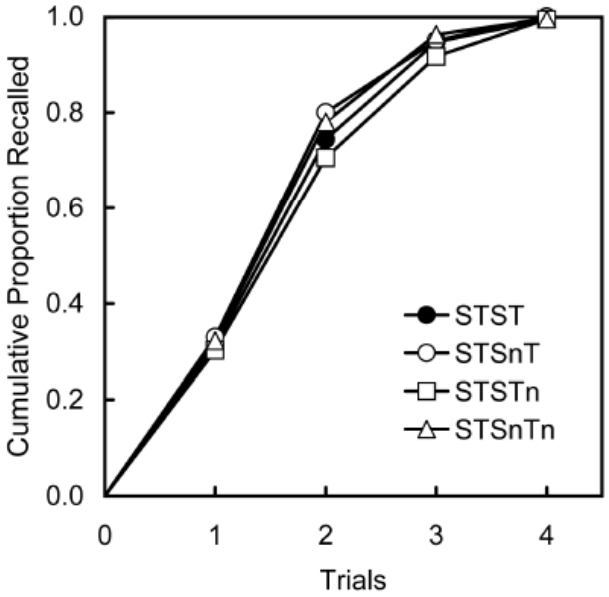
Les étudiants savent-ils qu'il vaut mieux se tester?

Karpicke, J. D., & Roediger, H. L., 3rd (2008). The critical importance of retrieval for learning. *Science, 319, 966-968.*

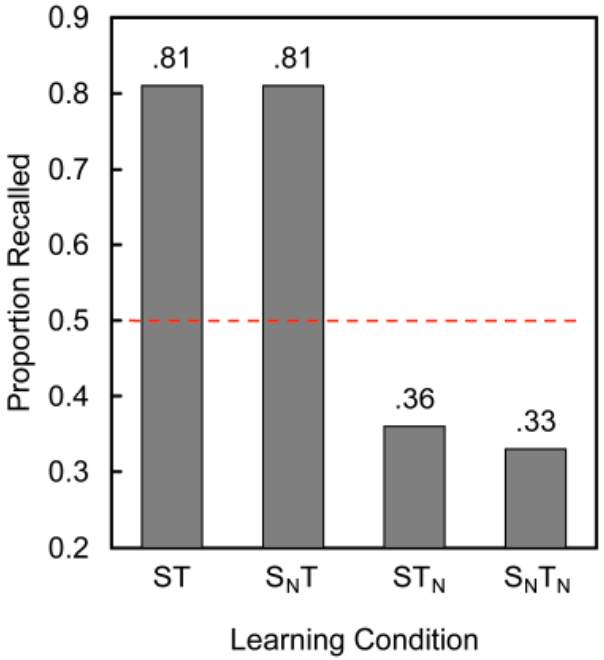
Apprentissage de traductions de 40 mots en swahili.
Quatre conditions:

Study all, test all	ST ST
Study all, test nonrecalled	ST ST_N
Study nonrecalled, test all	ST S_NT
Study nonrecalled, test nonrecalled	ST S_NT_N

Phase d'apprentissage: performances identiques



Prédictions métacognitives des étudiants: « je me souviendrai d'environ 20 mots [50%] » -- sans différence entre conditions



Et pourtant… c'est le test répété de tous les items qui compte pour la rétention une semaine plus tard!
Karpicke, Butler & Roediger (*Memory* 2009) montrent que les étudiants ne pensent pas à se retester régulièrement (illusion de compétence!)

L’importance de la métacognition pour l’éducation et l’auto-éducation

La représentation, par l'élève, des connaissances qu'il possède et de la façon dont il peut les améliorer est considéré, par certains pédagogues, comme un élément essentiel de l’éducation:
- comment as-tu fait pour comprendre?
- qu’est-ce que tu ne sais pas? Comment peux-tu trouver l’information pertinente?
- comment peux-tu faire pour apprendre mieux?

Cependant, l’apprenant ne connaît pas toujours bien son propre fonctionnement.
“Les étudiants ignorent les facteurs qui affectent leur propre apprentissage, ce qui a des implications importantes pour l’éducation.”
“Rendre les conditions d’apprentissage plus difficiles, ce qui oblige les étudiants à un surcroît d’engagement et d’effort cognitif, conduit souvent à une meilleure rétention.”
(Zaromb, Karpicke et Roediger, 2010)

Comment fonctionne le jugement métacognitif prospectif?

- Diverses théories:
  - jugements simplement fondés sur la récupération en mémoire?
  - Non: Reder (1987) demande aux participants, soit de répondre à une question simple, soit de dire s'ils savent (*feeling of knowing*). La récupération explicite de la réponse prend bien plus de temps que le jugement méta-cognitif.
  - Expérience de Son et Metcalfe (2005): temps de réaction très rapide pour savoir qu'on ne sait pas (catégorie de JOL la plus basse), alors que temps de récupération le plus long.
  - Utilisation d'**heuristiques** : des stratégies qui fournissent une solution rapide mais pas nécessairement optimale à un problème posé.
  - Le caractère plus ou moins approprié des indices utilisés peut expliquer, au moins en partie, pourquoi le jugement métacognitif apparaît souvent mal calibré (voir Gigerenzer et al., 1991).

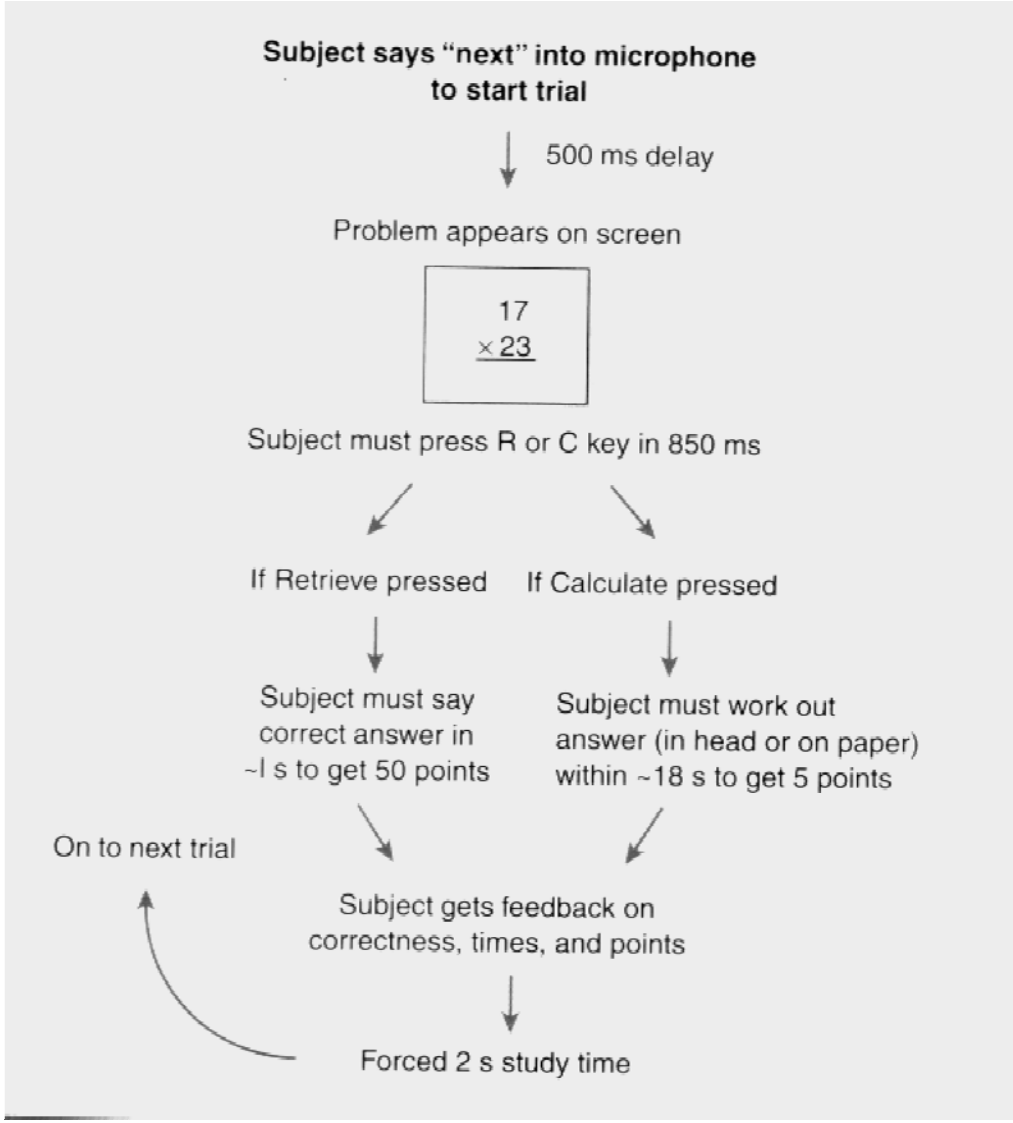
Le sentiment de savoir se fonde en partie sur une évaluation de la familiarité du problème

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(3), 435-451.*

Face à un problème arithmétique, les participants décident rapidement s'ils sont capables ou non de retrouver le résultat en mémoire.

Au cours de la phase d'apprentissage, la fréquence de présentation varie: certains problèmes sont présentés fréquemment, d'autres plus rarement.

S'ensuit une phase de test où les mêmes problèmes sont représentés, plus d'autres problèmes nouveaux.



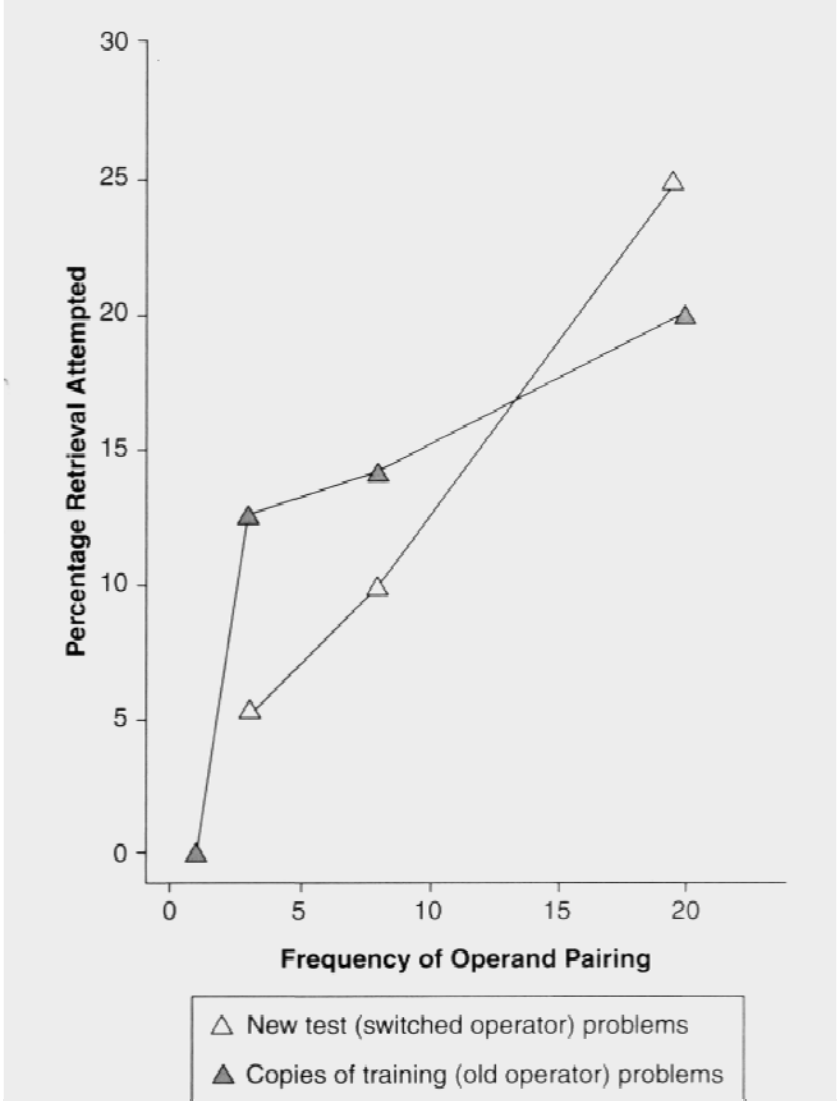
Le sentiment de savoir se fonde en partie sur une évaluation de la familiarité du problème

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(3), 435-451.*

Résultats:
Les participants jugent qu'ils peuvent récupérer l'information d'autant plus souvent que le problème a été présenté fréquemment (triangles pleins).
= métacognition correcte

Mais ils étendent également cette stratégie à des problèmes nouveaux mais qui ressemblent aux anciens (addition au lieu d'une multiplication):
-Pour les additions nouvelles qui ressemblent aux multiplications, ils tentent de retrouver le résultat en mémoire lors de 58% des essais.
- Alors qu'ils ne font cela que 10% du temps pour des multiplications entièrement nouvelles.

Conclusion: le sentiment de savoir est issu d'un sentiment de familiarité avec le problème posé, et non d'une récupération rapide de la réponse.



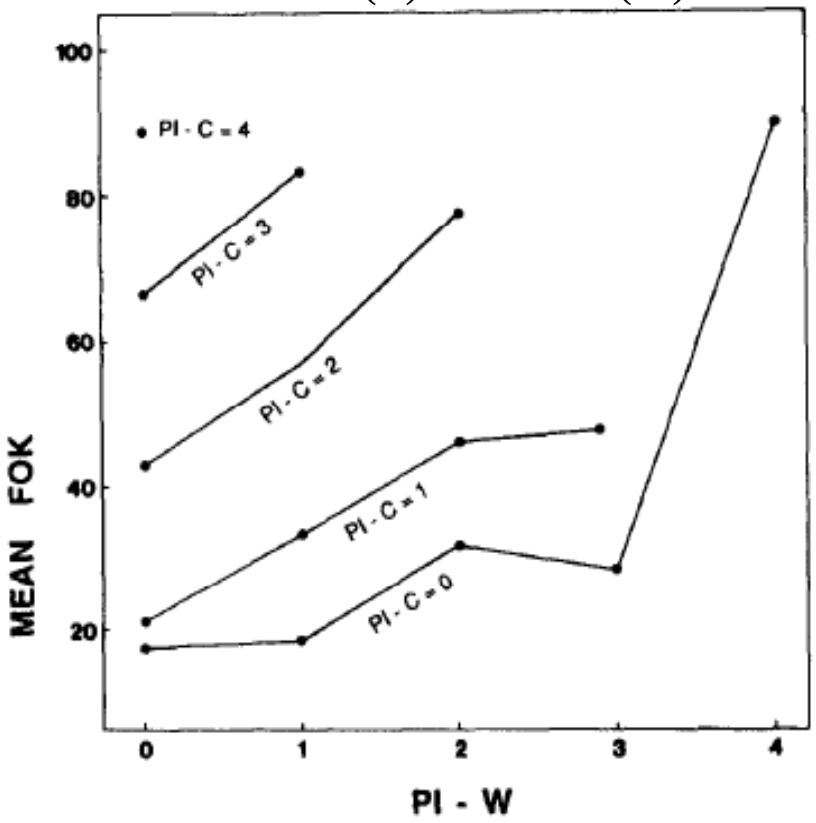
Le sentiment de savoir se fonde également sur l'accessibilité en mémoire de fragments de souvenirs

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychol Rev, 100(4), 609-639.*

Lorsque nous ne parvenons pas à nous rappeler une information, nous avons parfois accès à des connaissances partielles.
Selon Koriat (1993), le sentiment de savoir augmente de façon monotone avec l'accessibilité de ces informations – qu'elles soient correctes ou non.
Expérience de Koriat (1993):
- étude d'une série de 4 consonnes (FKDR) pendant 1 seconde, puis distraction pendant 19 secondes.
- tentative de se souvenir d'un maximum de lettres (mais avec une pénalité pour les lettres incorrectes)
- jugement de « sentiment de savoir »
- jugement de reconnaissance parmi 8 chaînes possibles.

Résultat:
- le sentiment de savoir corrèle avec la reconnaissance, mais il est prédit aussi bien par les lettres correctes que par les lettres incorrectes

Feeling of knowing
en fonction du nombre de lettres correctes (C) ou fausses (W)



Une théorie générale du rapport verbal

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87(3), 215-251.*

Ericsson et Simon introduisent une classification des tâches introspectives qui distingue le *moment* du rapport verbal (immédiat ou différé), et le *type* de rapport (direct, avec recodage, ou sans relation).

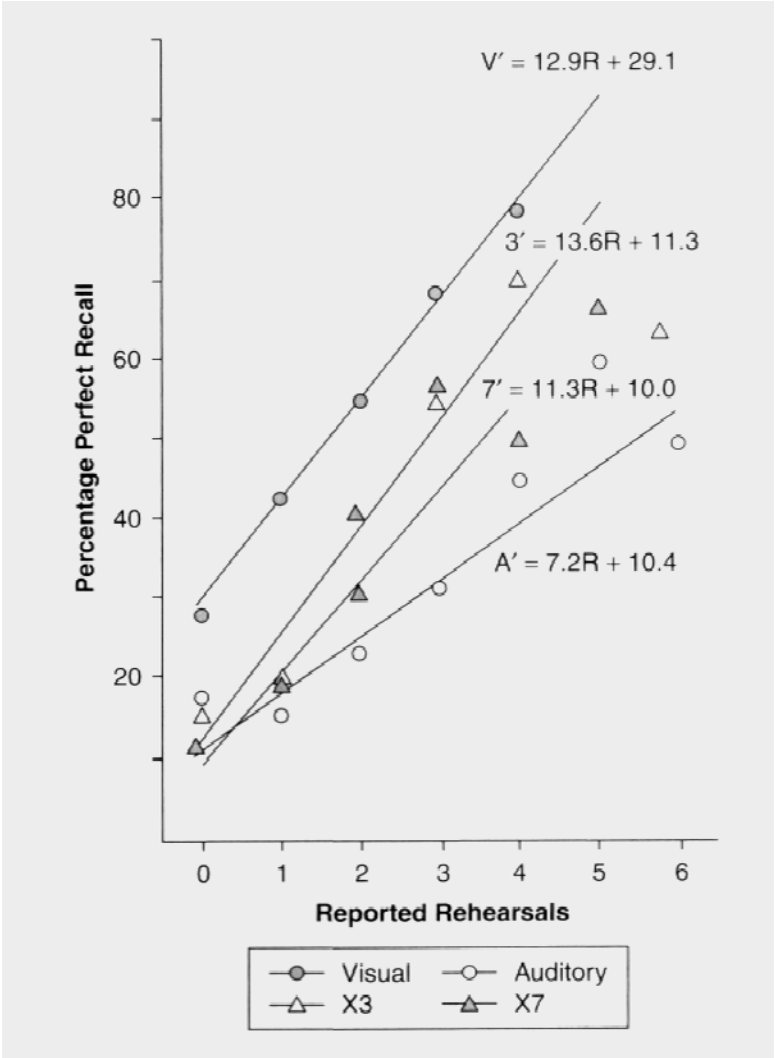
A Classification of Different Types of Verbalization Procedures as a Function of Time of Verbalization (Rows) and the Mapping From Heeded to Verbalized Information (Columns)

	Relation between heeded and verbalized information			
		Intermediate processing		
Time of verbalization	Direct one to one	Many to one	Unclear	No relation
While information is attended While information is still in short-term memory	Talk aloud Think aloud Concurrent probing	Intermediate inference and generative processes		
After the completion of the task-directed processes	Retrospective probing	Requests for general reports	Probing hypothetical states	Probing general states

Ce modèle prédit que l'introspection verbale doit être excellente dans un cas bien précis:
- lorsque le rapport est immédiat ou dans un délai très court
- et que l'on demande au sujet de rapporter une information présente dans sa mémoire de travail.

Un exemple de méta-cognition correcte: l'introspection du nombre d'opérations réalisées

Kroll, N. E. A., & Kellicut, M. H. (1972). Short-term recall as a function of covert rehearsal and of intervening task. *Journal of Verbal Learning and Verbal Behavior, 11, 196-204.*



Kroll et Kellicut présentent à leurs participants des triplets de lettres, et mesurent la capacité de rappel en mémoire après un délai avec distraction.

Les sujets rapportent combien de fois ils ont répété mentalement le triplet.

Quelle que soit la condition expérimentale, le pourcentage de triplets correctement rappelés est une fonction directe de l'introspection des sujets.

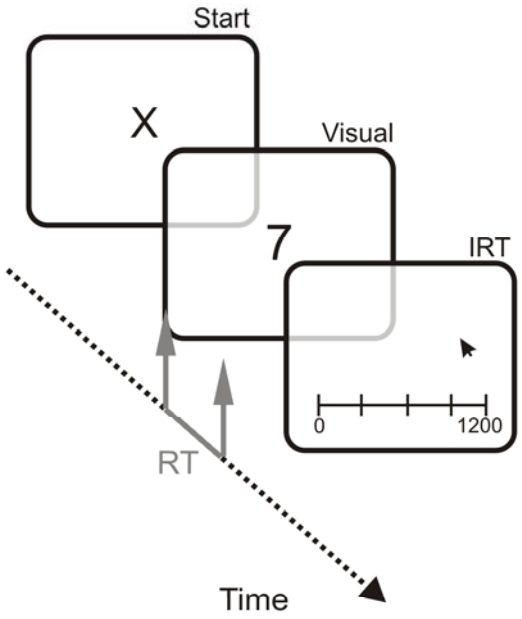
Plus impressionnant encore, l'introspection est un meilleur prédicteur que d'autres mesures objectives de la répétition mentale.

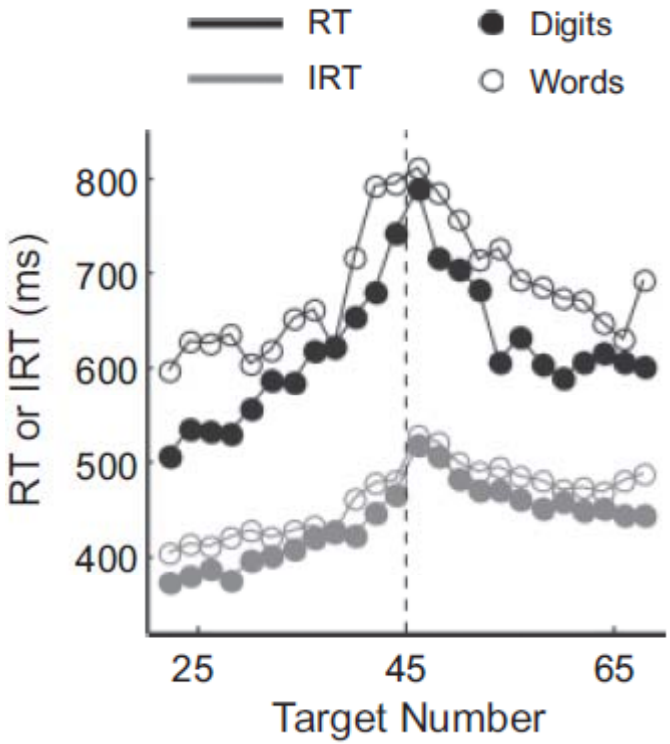
L’introspection du temps consacré à une tâche

(Corallo, Dehaene, Sackur & Sigman, *Psychological Science, 2008*)

Avons-nous une bonne « **introspection quantitative** » de la durée d’une étape de traitement mental? Après chaque essai, nous avons demandé aux sujets de dire combien de temps ils avaient mis à répondre.

Tâche simple:
Comparaison de deux nombres.
L’introspection est excellente

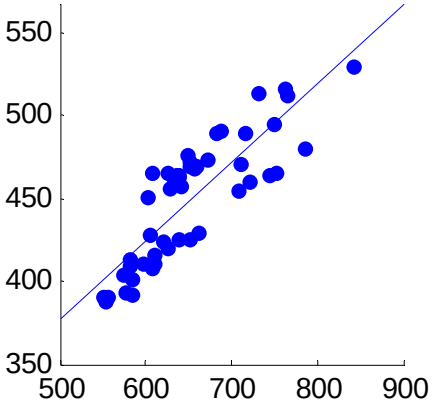




temps de réponse objectif

Introspection

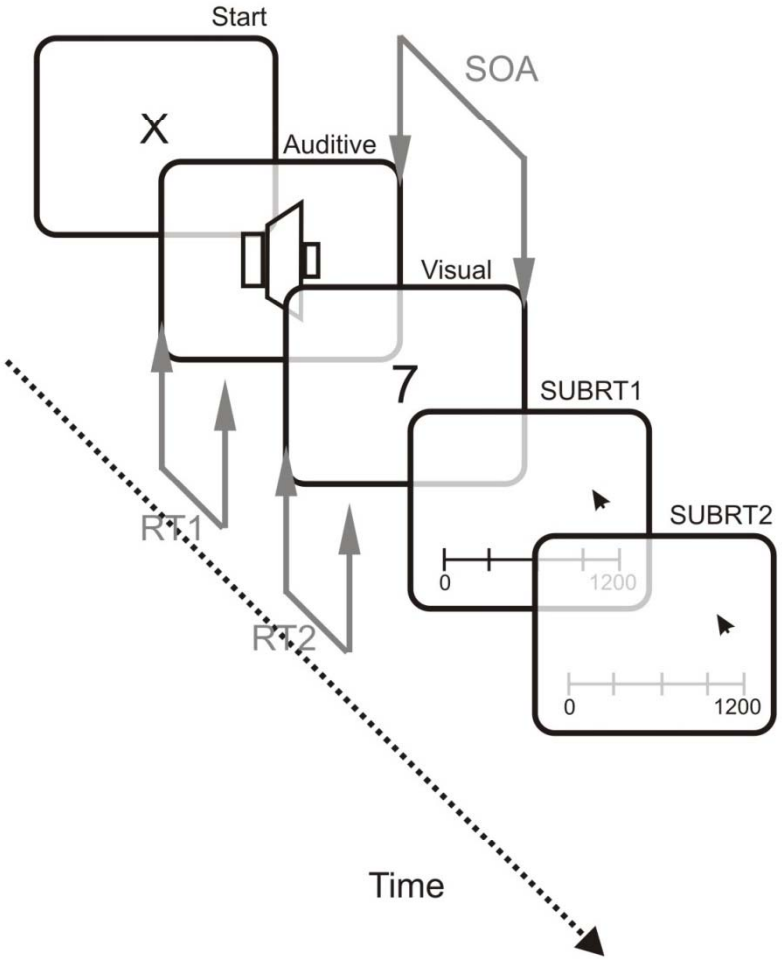
Corrélation de RT et IRT

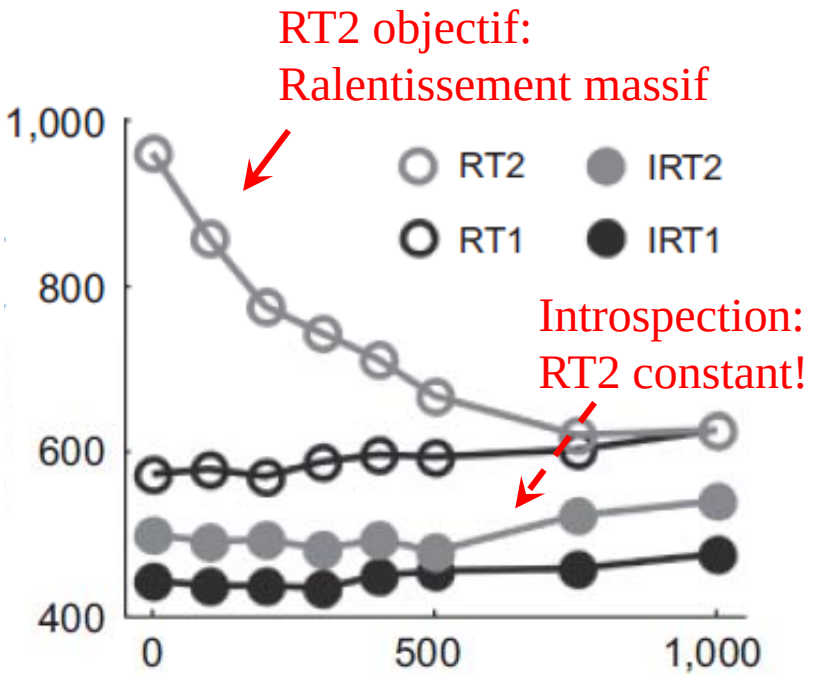


Echec de l'introspection lors de l'exécution d'une tâche double

(Corallo, Dehaene, Sackur & Sigman, *Psychological Science, 2008*)

Tâche double:
Jugement de sons, puis comparaison de nombres
L'introspection est mauvaise. Les sujets n'ont pas conscience du ralentissement de la seconde tâche.

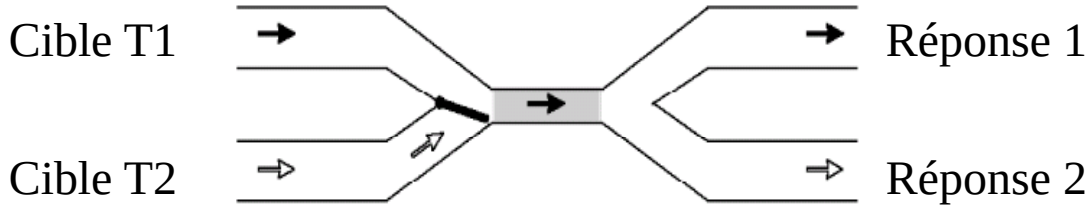




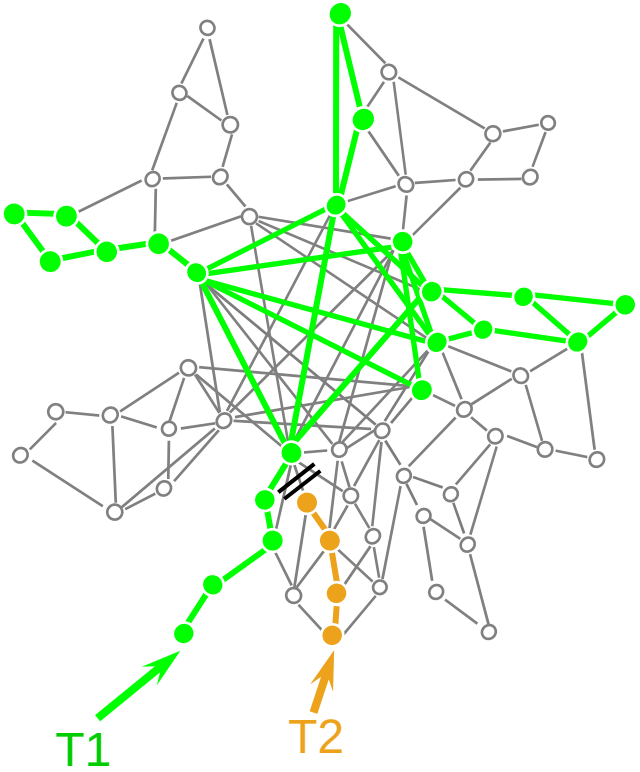
Délai entre T1 et T2

Introspection, espace de travail global, interférence centrale sont étroitement reliés

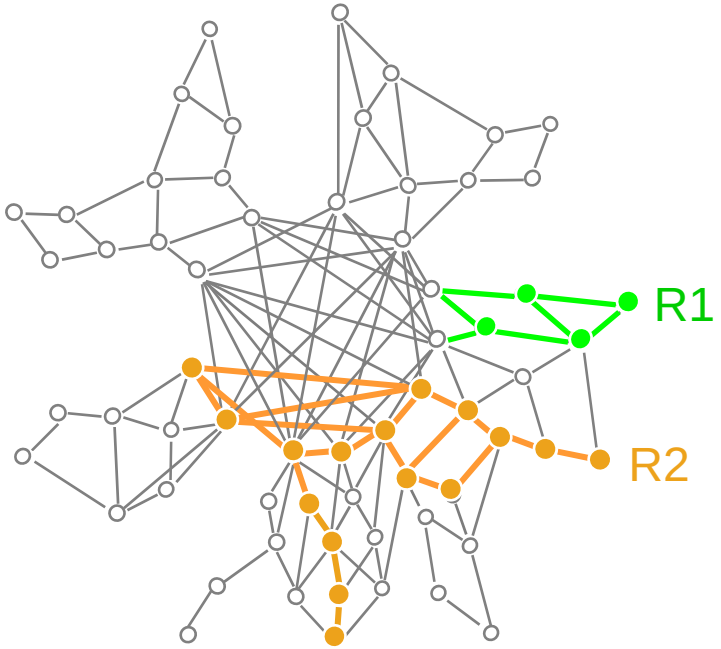
L'hypothèse du goulot d'étranglement central (Pashler, 1994)



Etat cérébral pour T1



Etat cérébral pour T2



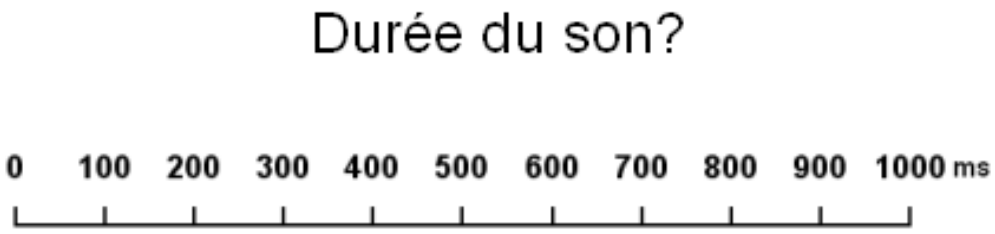
Sigman and Dehaene, PLOS Biology, 2005

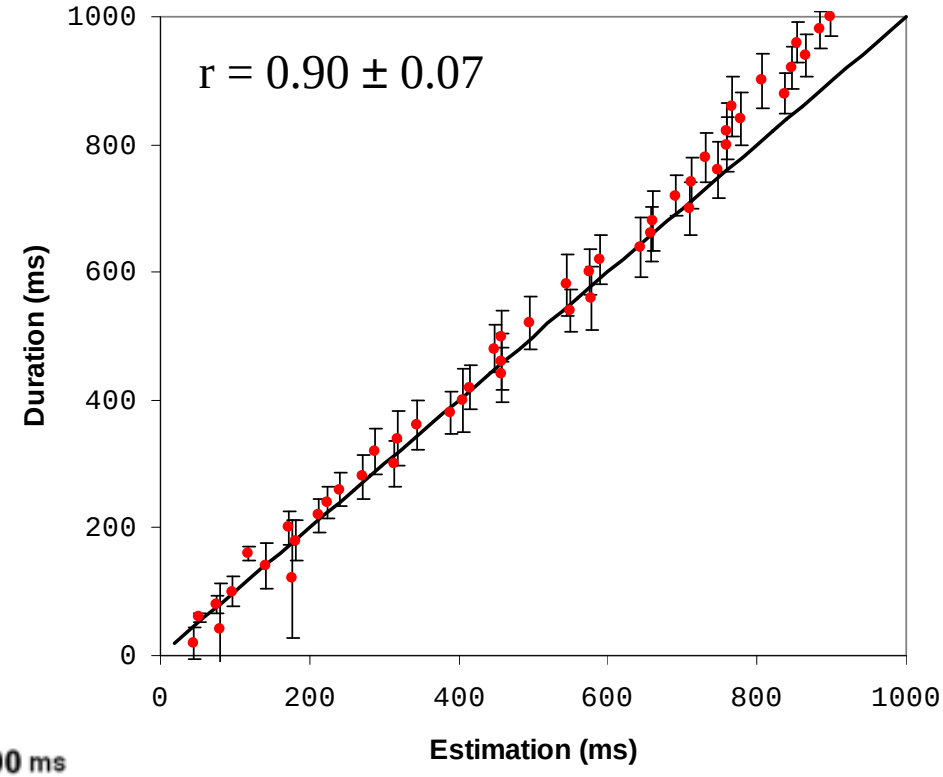
Une limite centrale de l'introspection

Marti, Sackur, Sigman et Dehaene, *Cognition*, 2010

Un premier contrôle:
Entraînement des sujets pour calibrer leurs réponses

- Estimation de la durée d'un son, avec feedback
- Durées: 20-1000 ms





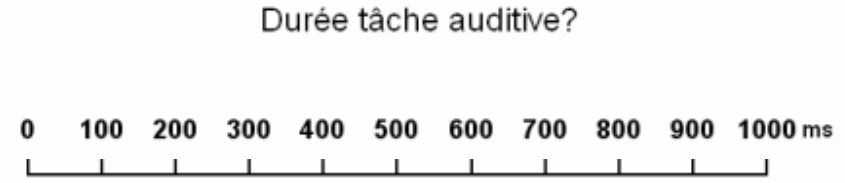
Une limite centrale de l'introspection

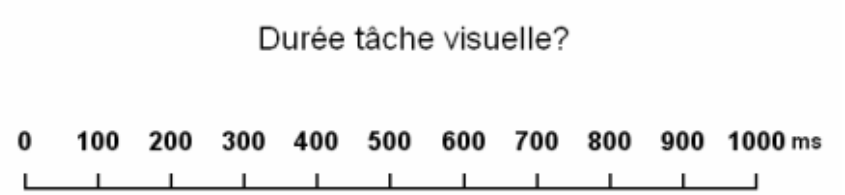
Marti, Sackur, Sigman et Dehaene, *Cognition*, 2010

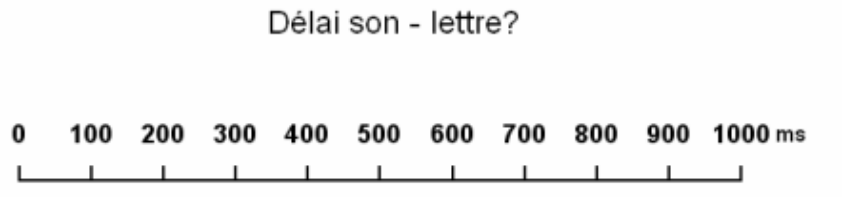
Deuxième contrôle:

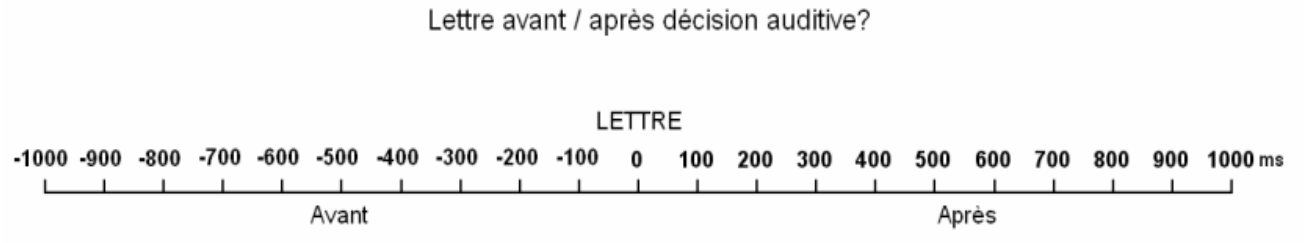
Pas moins de quatre questions précises sont posées au sujet

- Introspection de RT1: Combien de temps pour répondre au stimulus auditif?
- Introspection de RT2: Combien de temps pour répondre au stimulus visuel?
- Introspective du délai T1-T2: Combien de temps s'est écoulé entre le stimulus auditif et le stimulus visuel?
- Introspection du temps libre: le stimulus visuel est-il apparu avant ou après votre décision auditive? Combien de temps avant ou après?









Une limite centrale de l'introspection

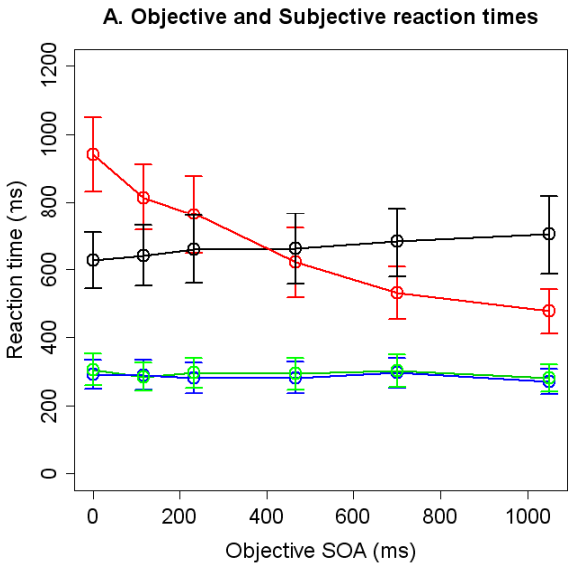
Marti, Sackur, Sigman et Dehaene, *Cognition*, 2010

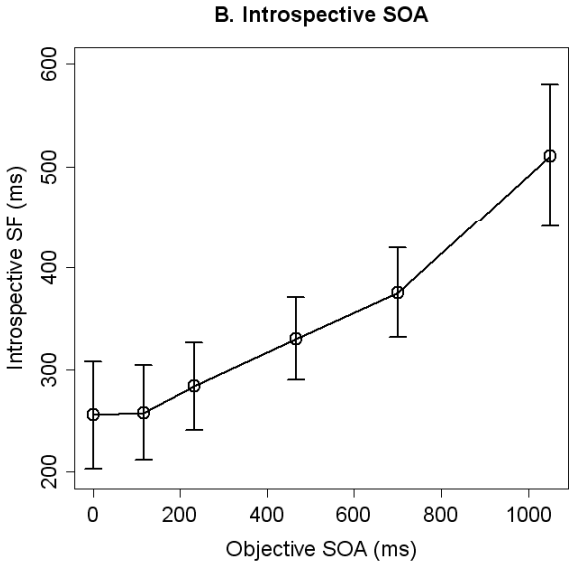
L'introspection de RT1 est excellente

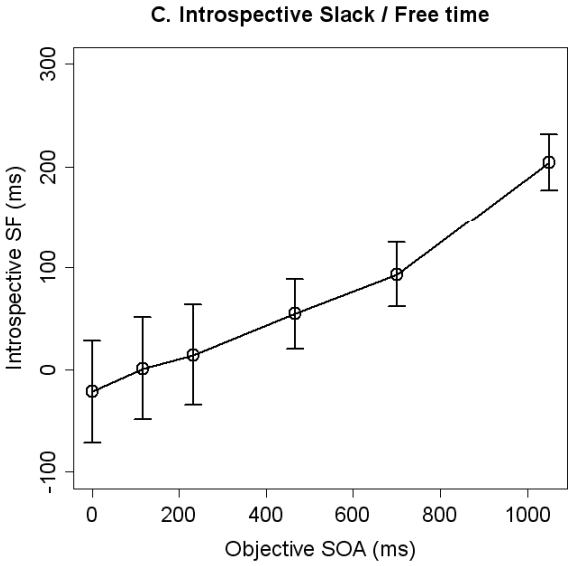
l'introspection de RT2 ne prend pas en compte le temps pris par T1

Les sujets surestiment le délai T1-T2 aux temps courts: Pendant qu'ils sont occupés par T1, ils ne savent pas quand T2 a été présenté.

Les sujets ont une bonne introspection du temps libre, (quand T2 arrive après la décision sur T1)
mais pas du temps d'attente (quand T2 doit attendre la fin de la décision sur T1)



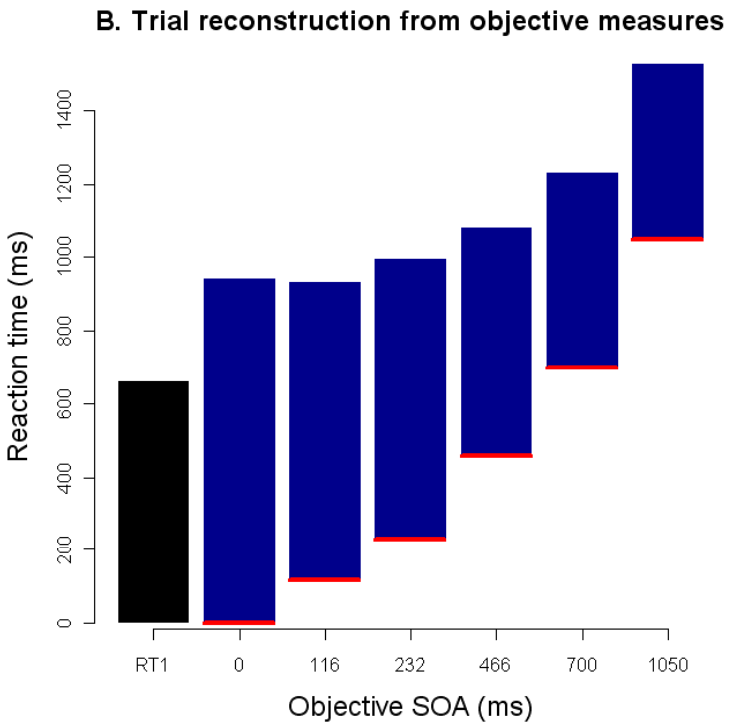




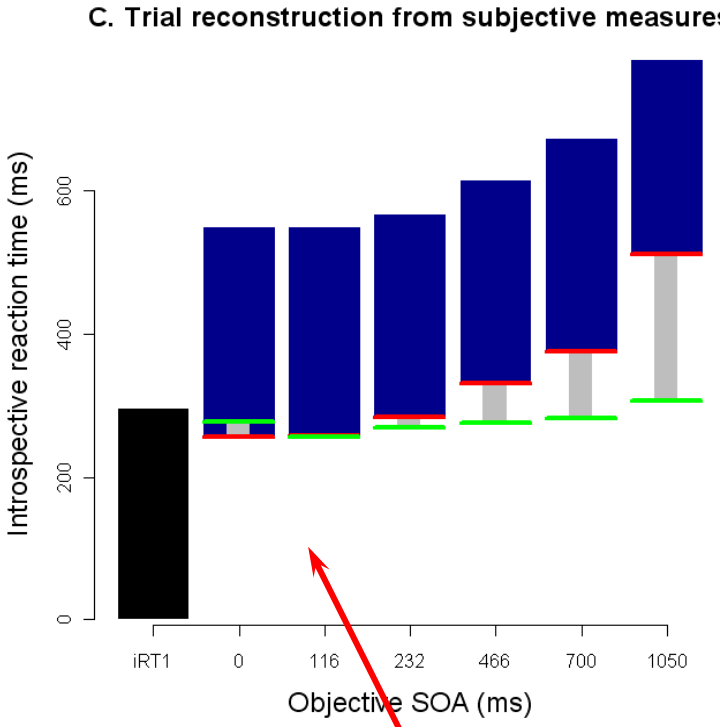
Une limite centrale de l'introspection

Marti, Sackur, Sigman et Dehaene, *Cognition*, 2010

Evénements objectifs *d'un essai*



Introspection subjective *d'un essai*



La tache aveugle de l'introspection!

Conclusions

Notre introspection est massivement limitée.

- Nous souffrons d'illusions métacognitives, par exemple lorsque nous croyons approcher de la solution d'un problème, ou bien lorsque nous pensons avoir suffisamment étudié une question.

- Seule l'expérimentation sur nous-mêmes nous permet d'évaluer nos connaissances

-Adultes, nous avons développé diverses heuristiques pour juger de notre mémoire:
1. familiarité du problème
2. accès à des connaissances partielles

-Il existe au moins un cas où notre capacité d'introspection est réelle:
Le contenu actuel ou récent de l'espace de travail neuronal conscient semble directement accessible à l'introspection.

- Malheureusement, de très nombreuses informations échappent à cet espace de travail

    d'une part parce qu'il est lent et sériel

    d'autre part parce qu'il n'a pas accès aux traitements non-conscients (informations non-attendues, transitoires, codées par des processeurs spécialisés ou par leurs connexions, etc; cf cours 2009)