Validació
i
Fiabilitat
de Proves
de Rendiment
d'Anglès

RECERCA I ESTUDIS

EN L'ÀMBIT DE

l'Avaluació dels aprenentatges

i l'Autoavaluació de Centres

PER AL

DEPARTAMENT D'ENSENYAMENT

GENERALITAT DE CATALUNYA

Autora: Rosario Outes Jiménez

DESITJO AGRAIR

AL DEPARTAMENT D'ENSENYAMENT DE LA GENERALITAT, I EN ESPECIAL ALS RESPONSABLES DE LA SUBDIRECCIÓ GENERAL DE FORMACIÓ PERMANENT I RECURSOS PEDAGÒGICS, L'OPORTUNITAT DE PODER DUR A TERME AQUEST TREBALL.

A L'INSPECTOR JESÚS RUL, L'AJUT I EL BON CRITERI DEL QUAL M'HAN RECOLZAT AL LLARG DE LA MEVA RECERCA.

A LES ESCOLES I ALS SEUS PROFESSORS, QUE M'HAN REBUT I HAN COL·LABORAT DESINTERESSADAMENT AMB EL DISSENY I L'APLICACIÓ DE LES PROVES.

I ALS ALUMNES, QUE HAN PARTICIPAT ACTIVAMENT I DE BON GRAT.

Introducció

Aquesta recerca està centrada en el desenvolupament d'una prova d'anglès del nivell de primer EOI com a instrument de mesura del rendiment en tasques de competència escrita, elaborada amb uns criteris específics i objectius de **fiabilitat i validesa**. En la seva posterior administració sobre una mostra representativa d'alumnes en tres escoles de Barcelona. I, per últim, en l'anàlisi i la descripció estadística dels resultats mitjançant els programes SPSS, Minitab i Excel.

Considerant les qualitats específiques que determinen la utilitat de la prova, és essencial prendre un punt de vista sistemàtic, considerant-la com a part d'un context educatiu més gran i amb uns objectius de competències més amples. Parlem doncs de l'ús del test escrit en un sistema d'ensenyament, el qual inclou molts i molt diversos materials d'ensenyament , d'avaluació i d'activitats d'aprenentatge.

La diferència principal entre aquesta prova o les proves en general i altres components del programa d'ensenyament és el seu propòsit. Mentre que el propòsit primari d'altres components és proporcionar aprenentatge, el propòsit primari dels tests és mesurar. Encara que els tests poden servir per a propòsits pedagògics, aquest no és el seu propòsit primari.

Les tasques del sistema educatiu a les escoles d'idiomes comparteixen algunes característiques amb les tasques de les proves, com ara l'autenticitat, la interactivitat i l'efecte (washback).

Dues de les qualitats, però: **fiabilitat i validesa** són crítiques pels tests, i són de vegades referides com a qualitats de mesura essencials (Palmer i Bachman, 1996). Això és perquè aquestes són les qualitats que proporcionen la gran justificació per usar les puntuacions del test –números- com a base per a fer inferències o prendre decisions.

Són els principis bàsics que informen una prova amb criteris de *quality & fairness* (qualitat i justícia/imparcialitat). (*Standards of the Educational Testing Service.* Princeton, NJ).

Fiabilitat i validesa han estat tradicionalment descrites com a característiques més o menys independents. Això ha portat a alguns examinadors del llengües a posicions extremes, com ara l'idea que aplicant amb rigor una qualitat ens dóna la pèrdua virtual de l'altra. A vegades, s'ha dit que les qualitats de fiabilitat i validesa són necessàriament en conflicte, o que no és possible dissenyar tasques de test que siguin vàlides i fiables alhora.

És més raonable admetre què malgrat hi hagi una tensió entre les qualitats del test, això no significa que s'hagi d'abandonar cap. En comptes d'emfatitzar la tensió entre les diferents qualitats, els desenvolupadors del test necessitem reconèixer la seva complementarietat i la necessitat de maximitzar-les totes dues.

Quan incrementem la fiabilitat de les nostres mesures, estem alhora satisfent un condició necessària per a la validesa: per que el resultat del test sigui vàlid, aquest ha de ser fiable (de fet, els coeficients de correlació interna son base de estudis de fiabilitat i de validesa, segons diferents autors).

Hi ha (de cada vegada menys, això és cert) un cert maniqueísme, una tendència a segregar els aspectes més aviat qualitatius de l'anàlisi de validesa dels estudis de caire quantitatiu que es fan servir per treure els coeficients de fiabilitat. En aquest estudi es reflecteix que ambdues qualitats poden ser alhora millor enteses reconeixent-les com a aspectes d'un tret comú en la medició que cooperen. Identificant, estimant, i controlant els efectes dels factors que afecten als resultats del test.

La investigació de la fiabilitat té a veure amb la pregunta, "Quant de la nota obtinguda en un test per part de un individu és degut a la mesura de l'error, o a factors que no són de l'habilitat del llenguatge?" i minimitzar els efectes d'aquests factors en els resultats del test.

La validesa, per altra banda, té a veure amb la pregunta "Quant dels resultats del test per part de l'individu és degut a les habilitats del llenguatge que volem mesurar?" i maximitzar els efectes d'aquestes habilitats en els resultats del test.

Aleshores, els trets de fiabilitat i validesa poden ser vistos com a dos objectius que es conjuguen en dissenyar, desenvolupar i analitzar els tests per a: (1)

minimitzar els efectes de la mesura de l'error, i (2) maximitzar els efectes de l'habilitat del llenguatge que volem mesurar.

I aquests estudis es connecten i es retroalimenten des de l'inici: en l'elaboració de la prova, de forma exploratòria, fins al moment de l'anàlisi final, de forma confirmatòria.

Marc Teòric

Validesa

Una prova és vàlida quan és una eina de mesura adequada per mesurar allò que ens proposem mesurar (Henning, 1987).

La validesa de una prova fa referència a la significació i a l'adequació de les interpretacions que fem sobre la base de les puntuacions del test. Quan interpretem els resultats dels tests d'idiomes com a indicadors de l'habilitat comunicativa dels examinats, una qüestió crucial és fins a quin punt podem justificar aquestes interpretacions.

La implicació clara d'aquesta pregunta és que, com a professors, hem de ser capaços de proporcionar una justificació apropiada per a cada interpretació que fem del resultat d'un test donat. Això és, necessitem demostrar, o justificar, la validesa de les interpretacions que fem dels resultats del test, i no simplement afirmar que aquests són vàlids.

Per justificar una interpretació determinada dels resultats, necessitem proporcionar l'evidència que el resultat del test reflecteix les àrees de l'habilitat i coneixement del llenguatge que volem mesurar, i poca cosa més.

Les evidències es poden registrar de formes molt diferents. Per a les proves aplicades en aquesta recerca he aplicat uns criteris de validesa que es podrien agrupar sota els epígrafs de : validesa racional, validesa empírica i validesa del constructe (Alderson, Clapham i Wall, 1996).

Malgrat això, la validesa és un concepte unívoc: el grau de validesa es refereix al grau de certitud que aquestes evidències donen per què les considerem una base contrastada de les inferències que fem de les puntuacions obtingudes en una prova. I, més aviat, hauríem de parlar de la validesa de les inferències, i no pas de les puntuacions.

Aquesta estudi assimila de bon començament, la **taula d'anàlisis factorial** elaborada per Jesús Rul (2000), que constitueix una anàlisi de:

- validesa racional o dels continguts
- validesa de constructe o de les capacitats
- validesa empírica o de contrast: amb l'avaluació continuada i amb un estudi estadístic de correlacions de les parts.

Fiabilitat

En aquest estudi no hi ha un propòsit de proporcionar una guia pràctica per conduir una anàlisi de fiabilitat. En comptes d'això, s'argumenta la racionalitat i la lògica que fonamenta el concepte de fiabilitat, com a base en el desenvolupament, l'ús i l'interpretació dels tests de la llengua, i es descriu una prova específica, a la qual se l'aplica els criteris de fiabilitat i validesa que ara s'exposen.

Avui en dia comptem amb aplicacions informàtiques molt sofisticades que calculen els coeficients i dissenyen gràfics il·lustratius automàticament. Malgrat això, la comprensió raonada i l'operació manual de les fórmules és una tasca feixuga però absolutament indispensable per assolir un coneixement rigorós de la lògica que sustenta l'anàlisi empírica.

Un assumpte fonamental en el desenvolupament i en l'ús dels tests d'anglès és identificar les fonts potencials d'error en una mesura donada de l'habilitat del llenguatge i minimitzar l'efecte d'aquests factors en aquesta mesura. Hem de tenir presents els errors de mesura, o la no fiabilitat, perquè sabem que els resultats dels tests es veuen afectats per altres variables que no són les habilitats que volem mesurar. Per exemple, podem pensar en factors tals com a una malaltia, fatiga, manca d'interès o motivació, etc. que poden afectar la realització dels tests per part dels individus, però que no són generalment associats amb l'habilitat de comunicar-se en la llengua meta, i per tant característiques que no volem mesurar en les nostres proves. Aquestes, però, són algunes de les fonts més òbvies de l'error de medició.

A més de elements com aquests, els quals són extensament asistemàtics i, per tant impredictibles, les característiques del mètode de les tasques del test o facetes són fonts potencials d'error que poden ser igualment perjudicials per a la mesura acurada de les habilitats del llenguatge. I són precisament aquestes fonts d'error les que es poden prevenir o corregir mitjançant una anàlisi empírica. Quan minimitzem els efectes d'aquests diversos factors, minimitzem la mesura d'error i s'augmenta la fiabilitat.

En altres paraules, quant menys afectin aquests factors a la puntuació dels tests, millor serà l'efecte relatiu de les habilitats del llenguatge que volem mesurar, i per tant, la fiabilitat dels resultats dels tests del llenguatge.
Hi ha, doncs, alguns fets que assumim com a hipòtesi de treball:

- Fiabilitat i validesa es complementen per què si ens preguntem: "Quant dels resultats obtinguts en una prova per un individu és degut a l'habilitat del llenguatge que volem mesurar (com a oposició a les fonts d'error)?" podem veure que això també és essencial per l'examen de la validesa.
- el coeficient de fiabilitat i l'error de mesura de les proves (SEM) està generalment subestimat i gairebé mai no es té en compte. (James H. Mc Millan, 2000).
- el error de mesura i l'índex de fiabilitat són indicadors de qualitat que afecten les puntuacions i no pas les proves.

La investigació de la fiabilitat engloba per una banda, l'anàlisi lògica, i per l'altra, la recerca empírica; hem d'identificar les fonts d'error i estimar la magnitud dels seus efectes en els resultats del test. Per poder identificar les fonts d'error, necessitem distingir els efectes de les habilitats del llenguatge que volem mesurar dels efectes d'altres factors, i aquest és un problema complex.
Això és, per una banda, degut a la interacció entre els components de l'habilitat del llenguatge i les característiques del mètode del test. Pot ser difícil marcar una distinció clara entre l'habilitat per mesurar i les facetes del mètode. Identificar les fonts d'error és també complicat pels efectes d'altres característiques dels individus, com poden ser l'edat , l'estil cognitiu i l'experiència prèvia, que poden resultar també difícils de distingir-les dels efectes de les habilitats en la llengua que volem mesurar.
Mentre que les recerques han proporcionat aprofundiments en les relacions entre els factors com aquests i l'adquisició del llenguatge, els seus efectes en la realització de tests han començat recentment a investigar-se. Identificar

clarament i distingir els diversos factors que siguin fonts potencials d'error de la mesura en un test donat és crucial per a la investigació de la fiabilitat.
Estimar la magnitud dels efectes de diferents fonts d'error, un cop aquestes fonts han estat identificades, és un problema de recerca empírica, i és un aspecte bàsic de la teoria de la mesura. En aquesta recerca les estimacions estadístiques s'han fet a partir de les bases de dades obtingudes i arxivades amb format *.spss (SPSS), *.mtw (MINITAB) i *.xls (EXCEL).
Com que hi ha moltes fonts potencials d'error de mesura, nombrosos resultats han estat desenvolupats. I com que les fonts d'error poden implicar interaccions complexes, i afectar a diferents individus de forma diferent, els procediments requerits per una estimació adequada de la fiabilitat en els resultats d'un test són necessàriament complexos i requereixen uns coneixements tècnics mínims.
Els especialistes en la mesura coincideixen que l'examen de la fiabilitat depèn de la nostra habilitat per distingir entre els efectes (en els resultats dels tests) de les habilitats que volem mesurar i els efectes d'altres factors.
Els factors que causen que els resultats del test variïn d'un individu a un altre inclouen característiques *constants* generals i específiques, característiques *temporals* generals i específiques, i factors sistemàtics i de probabilitat relacionats amb el sentit del test i els resultats.
Els resultats de la prova poden veure's afectats per altres factors que no són específics de la competència en la llengua meta. Aquests poden ser agrupats en les següents grans categories:
1) les facetes del mètode del test:
Les facetes del mètode del test són sistemàtiques en el sentit que són uniformes d'un test a un altre. Això és, si la faceta o característica del métode del format de l'input és de elecció múltiple, això no variarà, tant si el test es donat al matí o a la tarda, o si el test es destina a mesurar la competència gramatical o lèxica.

2) els atributs de l'individu que realitza el test, que no són considerats part de les habilitats de la llengua que volem mesurar:

Els atributs dels individus que no estan relacionats amb la seva competència de la llengua (com ara l'estil cognitiu, el coneixement de àrees de contingut particulars, l'edat, etc.) són també sistemàtics en el sentit que afecten regularment a la realització d'un test.

3) factors aleatoris que són altament impredictibles i temporals:
Per altra banda, hi ha els factors asistemàtics, o aleatoris. Aquests inclouen condicions com ara el estat emocional i d'alerta mental o diferències en les condiciones físiques en l'administració de la prova.
Per això, qualsevol inferència que fem sobre el nivell de l'habilitat comunicativa en base a la seva puntuació al test serà un error fins a cert punt si:

1. la puntuació del test de l'individu és afectada per les facetes del mètode del test,
2. volem mesurar altres atributs que no siguin les habilitats del llenguatge,
3. hi ha influència dels factors aleatoris.

Els resultats dels efectes de tots aquests factors és que sigui on sigui on els individus facin un test de la llengua meta, no tots actuaran de la mateixa manera, i per tant els seus resultats variaran. Si considerem la quantitat total de la variació en els resultats dels diferents individus en un test donat, podem pensar en proporcions diferents d'aquesta variació degudes a diferents factors discutits anteriorment. Això és perquè els diferents factors afectaran a diferents individus diferentment.
Fonamentalment els resultats varien pel nivell de les seves habilitats lingüístiques, però també variaran en el punts en què són afectats pels diferents mètodes de test (alguns individus, per exemple, poden fer-ho molt bé en un tasca d'omplir espais buits i fer-ho malament en un diàleg obert). O es pot veure el cas en que els membres d'un grup d'edat poden fer-ho millor o pitjor en un test donat que els membres d'altres grups. O, alguns individus que facin el test poden estar més descansats mentre que altres poden haver-hi estat malalts i conseqüentment no són capaços de realitzar el test segons el seu nivell de competència real.

El interès primordial en l'ús dels tests del llenguatge és fer inferències sobre la competència comunicativa de l'individu. És vital que en el disseny, el desenvolupament i l'interpretació dels resultats de les proves es minimitzin els efectes del mètode de test, dels atributs personals i els factors aleatoris (fonts d'error en la mesura) que no reflecteixen aquesta competència.

Quan investiguem la fiabilitat, és essencial tenir en compte la distinció entre **habilitats no observables**, per una banda, i les **puntuacions observades** del test, per l'altra. Les habilitats del llenguatge que ens interessa mesurar són abstractes, i per tant no les podem observar

directament, o saber, en sentit absolut, la puntuació "vertadera" de l'individu. Això solament pot ser estimat per la puntuació realment obtinguda en un test donat en aquesta habilitat. Per tant, qualsevol intent d'estimar la fiabilitat d´un conjunt de puntuacions d'un test ha d'estar basat en un model que especifiqui les relacions hipotetitzades entre l'habilitat mesurada i els resultats observats al test.

1. Precisió de Conceptes

Abans de presentar els fonaments teòrics de l'anàlisi empírica de fiabilitat que s'ha fet servir en aquest estudi, hauríem de aclarir tres conceptes estadístics bàsics: **la mitjana, la variància i la desviació estàndar**.

a. **La mitjana**, aquí simbolitzada amb la μ o M és la mitjana aritmètica dels resultats d'un grup donat de alumnes.

μ= $\frac{\sum X}{N}$ és la suma dels valors dividida pel nº de dades

b. **La variància**, simbolitzada amb s^2, és un estadístic que caracteritza quants resultats dels individus varien de la mitjana del grup. La variància és igual al quadrat de la desviació estàndard, s o s .

c. I **la desviació** és l'arrel quadrat de la mitjana dels quadrats de les desviacions de cada dada (x_i) respecte la mitjana.

$$\sigma = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{N}}$$

Aquesta és la fórmula de la desviació de la població, en aquesta recerca ens referirem a la desviació corregida o de la mostra (N-1) per la qual cosa només cal posar N-1 en el denominador de l'arrel quadrat.

Subíndexs tal com "x, t" ó "e " seran utilitzats per indicar tipus específics de variància. El símbol s^2_x per exemple, serà utilitzat per referir-nos a la variància en les puntuacions dels tests observats.

2. La teoria clàssica de la mesura

La teoria clàssica de la mesura de la puntuació vertadera (CTS) consisteix en un conjunt de suposicions sobre les relacions entre els resultats del test vertader o observat i els factors que afecten aquests resultats.

® La primera suposició d'aquest model afirma que un resultat observat en un test inclou dos factors o components: **un *resultat vertader* que és degut al nivell de competència de l'individu i un *resultat d'error*, que és degut a altres factors que no són de l'habilitat testada**. Aquesta suposició pot ser representada a la fórmula:

$$x = x_v + x_e$$

"x" és el resultat observat,
"x_v" és la puntuació vertadera,
"x_e" el resultat d'error.

Similarment, podem caracteritzar la variància d'un conjunt de puntuacions de test consistint en dos elements:

$$\mathbf{s^2_x = s^2_v + s^2_e}$$

"s^2_x" és la variància del resultat observat,
"s^2_t" és el component de variància del resultat vertader,
"s^2_e" és el component de la variància del resultat d'error.

® Una segona suposició té a veure amb la relació entre els resultats vertaders i d'errors. Essencialment, aquestes suposicions afirmen que els resultats d'error són asistemàtics, o aleatoris, i no són correlatius amb els resultats

vertaders. Sense aquestes suposicions no seria possible distingir les puntuacions vertaderes de les puntuacions d'error. Aquestes suposicions constitueixen la **definició de l'error** de mesura del model CTS com a aquella **variació en un conjunt de resultats de test que és asistemàtica o aleatòria**.

En resum, el model CTS de mesura defineix dues fonts de variància en un conjunt de resultats de un test: la variància del resultat vertader, la qual es deguda a les diferències en l'habilitat dels individus testats, i la variància de l'error de mesura, la qual és asistemàtica, o aleatòria.

Dins el model CTS, hi ha tres tipus de estimació de fiabilitat, cadascuna enfoca de forma diferent l'anàlisi de les fonts d'error.

- Les estimacions de la *consistència interna* es refereixen primordialment a les fonts d'error i als resultats dins del mateix test,
- Les estimacions d'*equivalència* (o de proves paral·leles) que es calculen tenint en compte les puntuacions de formes alternatives d'un test.
- Les estimacions de l'*estabilitat* (test-retest) indiquen quant de consistents són els resultats del test a través del temps. Aquestes estimacions no s'han pogut dur a terme perquè el marge de temps no ho va permetre.

Les estimacions de la fiabilitat que proporcionen aquestes aproximacions s'anomenen *coeficients de fiabilitat*.

3. Tests paral·lels

El model d'estimació de fiabilitat per l'equivalència CTS és el dels *tests paral·lels*. Per què dos tests es considerin paral·lels, assumim que són mesures de la mateixa habilitat, això és, que el resultat vertader de l'individu en un test serà el mateix que el seu resultat vertader en l'altre. Però, com que realment mai sabem els resultats vertaders per a un test donat, com podríem saber si els dos tests són paral·lels? La resposta a aquesta pregunta ens la proporciona el fet de que els tests paral·lels són definits en la teoria clàssica de mesura de la següent manera:
Dos tests són paral·lels si, per cada grup de persones que fan els dos tests, el resultat vertader d'un test és igual que el resultat vertader de l'altre, i les variances de l'error dels dos tests són iguals. D'aquesta definició podem derivar una definició operacional: els tests paral·lels són dos tests de la mateixa habilitat que tenen la mateixa mitjana i variància.
Per tant dos tests, X i X´ , donant els resultats *x* i *x´*, poden ser considerats paral·lels si es donen les condicions següents:

$$\mu = \mu'$$
$$s^2_x = s^2_{x'}$$

A la pràctica, tractem dos tests com si fossin paral·lels si les diferències entre les seves mitjanes i variances no són estadísticament significatives.

Les definicions de resultat vertader i de la variància del resultat d'error donades a dalt són abstractes, en el sentit que no podem certament observar el resultat vertader i d'error en un test donat.
Mai sabem quins són els resultats vertaders o d'errors, no podem *saber* la fiabilitat dels resultats observats. Per poder *estimar* la fiabilitat dels resultats observats, per tant, hem de definir la fiabilitat operacionalment d'una manera

que depengui únicament dels resultats observats (Mc Millan, 2000), i per això utilitzem la definició operacional dels tests paral·lels.

Es suposa que els errors de mesura dels dos tests són aleatoris, i no seran correlatius. A causa de l'efecte dels resultats d'error aleatoris, la correlació entre els resultats dels tests paral·lels serà menys que perfecta. Quant menys efecte de resultats d'error hi hagi, més correlació hi haurà en els tests paral·lels. Per tant, si els resultats observats en dos tests paral·lels són altament correlatius, això indica que els efectes dels resultats d'error són mínims, i que poden ser considerats indicadors fiables de l'habilitat que es mesura. No cal dir que l'habilitat mesurada ha d'estar la mateixa.

D'això podem inferir una definició de la fiabilitat com la correlació entre els resultats observats en dos tests paral·lels, que podem simbolitzar com $r_{xx'}$. Aquesta és la definició que ens proporciona la base per a totes les estimacions de la fiabilitat a la teoria CTS.

Una qüestió fonamental per a aquesta definició operacional de la fiabilitat és la suposició de que els resultats observats a ambdós tests són **experimentalment independents**. Això és, l'actuació per part de l'individu al segon test no hauria de dependre de com ell/ella ho ha fet en el primer. Si l'actuació de l'individu en el primer test influencia la seva actuació en el segon, llavors no podem fer servir la correlació entre els resultats observats com a base del càlcul de fiabilitat.

A més, com que podem estimar la fiabilitat mitjançant el càlcul de la correlació entre els tests paral·lels, també podrem inferir una estimació de l'error de mesura.
Si els resultat observat d'un individu en un test es compon d'un resultat vertader i d'un resultat d'error, quant més gran sigui la quantitat de resultat vertader, menor serà la proporció de resultat d'error, i per tant més fiable serà la puntuació observada.

Una manera de definir la fiabilitat és la proporció de la variància del resultat observat que sigui la variància del resultat vertader, la qual pot ésser representada com es mostra a continuació:

$$r_{xx'} = s^2_v/s^2_x$$

on r_{xx} és el coeficient de fiabilitat

Com que les proporcions de la variància del resultat vertader i de la variància del resultat d'error sumen 1, podem expressar la proporció de la variància d'error com a el complement de la fiabilitat:

$s^2_v/s^2_x + s^2_e/s^2_x = 1$

$s^2_e/s^2_x = 1 - s^2_t/s^2_x$

$(s^2_t/s^2_x = r_{xx'})$

$$s^2_e/s^2_x = 1 - r_{xx'}$$

Això porta a la definició de la variància del resultat d'error en termes de variància del resultat total i fiabilitat

$$\mathbf{s^2_e = s^2_x (1- r_{xx'})}$$

Aquesta definició és particularment útil, ja que l'estimació de la variància d'error i al càlcul de l'error estàndard i és més directament aplicable a resultats individuals de tests que les estimacions de fiabilitat. Perquè els coeficients de fiabilitat s'apliquen a grups de resultats mentre que les estimacions de la variància d'error es poden aplicar a **resultats individuals**.
En resumen, la fiabilitat CTS es definida en termes de la variància del resultat vertader. Com que mai podem saber els resultats vertaders dels individus, només podrem saber quina és la fiabilitat a partir dels resultats observats. I a més a més, com que la fiabilitat serà una funció no solament del test, però de

l´actuació dels individus que fan el test, qualsevol estimació donada de fiabilitat basada en el model CTS és doncs un valor relatiu, i es limita a la mostra a la qual s'apliquen els càlculs. Es refereix als resultats del test, i no al test mateix. Per tant, hem d'estimar sempre la fiabilitat dels resultats de grups específics amb els quals volem utilitzar el test.

De tota manera, encara que l'estimació del coeficient de fiabilitat a través de l'aplicació de proves equivalents és molt il·lustrativa de la teoria Classical True Score (CTS), hem d'admetre que la construcció de tests paral·lels és una tasca molt i molt dura. Impossible segons alguns autors (Alderson, Clapham i Wall, 1995). Però és fonamental copsar l'abast d'aquest anàlisi per tal d'entendre les proves de consistència interna, que és l'eina que es va fer servir més sovint en aquesta recerca.

4. Consistència interna

La *consistència interna* té a veure amb quant de consistència o coherència hi ha en les actuacions dels individus en les diferents parts del test. Les inconsistències en l'actuació de les diferents parts del test poden ser causades per un nombre de factors.

Per exemple, l'actuació en les parts del test de comprensió lectora pot ser inconsistent si els fragments són de diferent complexitat sintàctica i lèxica O els ítems de un tasca d'elecció múltiple poden tenir un grau de dificultat molt diferent. Aleshores, els resultats no serien indicadors massa fiables de la competència lingüística del alumne.

Les estimacions de fiabilitat per la consistència interna més habituals són:

- Les que es basen en la divisió en meitats (Spearman-Brown i Guttman).
- Les que es fonamenten en la variància dels ítems (Kuder-Richardson i Cronbach's Alpha).

5. Divisió en meitats (Split half)

Un mètode d'anàlisi de la consistència interna d'un test és el mètode de la comparació entre les meitats de la prova i determina el punt fins el qual els resultats d'aquestes dues meitats són consistents una amb l'altra.
Fent això, estem tractant les meitats com si fossin tests paral·lels, i per tant hem de fer certes suposicions sobre l'equivalència de les dues meitats, específicament que aquestes tinguin les mateixes mitjanes i variances.
A més a més, hem d'assumir també que les dues meitats són **independents** una de l'altra. Això és, que l'actuació d'un alumne en una meitat no condiciona quant ell fa a l'altra. Així, doncs, la correlació serà causada pel fet que ambdues estan mesurant el mateix tret o habilitat, i no pel fet que l'actuació d'una de les parts depengui de l'actuació de l'altra..
Per interpretar la relació entre les meitats com a indicador de que les dues mesuren la mateixa habilitat, o que les dues tenen la mateixa variància en el resultat vertader, hem de ser capaços d'excloure la segona interpretació.
Hi ha moltes formes de dividir una prova. Potser la més fàcil és la de partir el test en dues meitats aleatòries, amb el mètode "impar-par", pel qual tots els ítems amb numeració impar són agrupats junts en una meitat i tots els ítems amb numeració par en una altra. Aquest mètode és aplicable en tests en els quals els ítems estan dissenyats per mesurar la mateixa habilitat i són independents. El exemple més clar d'aquest tipus de prova són els elecció múltiple o *multiple choice*.
En molts casos, no obstant, partir el test en meitats equivalents és problemàtic, ja que un mètode aleatori de partició de test no és aplicable. Considerem, per exemple, un test de múltiples possibilitats al qual diferents ítems són dissenyats per mesurar habilitats diferents, com podria ser el cas d'un test de nivell d'una classe que pretén mesurar diversos objectius, com ara un listening i una tasca d'expressió escrita. En aquest cas, un mètode aleatori (per exemple, impar-par) ens donarà meitats que no són equivalents en contingut, i s'hauria de dividir el test en dues meitats basant-se en el contingut dels ítems, en comptes de fer-ho aleatòriament, per assegurar-nos que són comparables. En aquest cas, particularment hauríem d'anar en compte per assegurar que les dues parts obtingudes són equivalents.

En qualsevol cas sempre que es pugui assumir que tots els ítems en el test mesuren la mateixa habilitat del llenguatge i que el fet que un individu contesti un ítem donat correctament no depèn de si ell ha contestat altres ítems correctament, aleshores estaríem justificats per usar el mètode impar-par per partir la prova.

Una vegada el test s'ha partit en meitats, es torna a puntuar, donant dos resultats –un per a cada meitat- per cada individu examinat. La fiabilitat estimada s'obté computant la correlació entre els dos grups de resultats. Això ens dóna una estimació de la consistència de les parts. En general, un test llarg serà més fiable que un test curt.

En el mètode de partir per la meitat, hem reduït el test a la meitat, i aleshores hem de corregir la correlació obtinguda per aquesta reducció de longitud.

Les estimacions de la correlació es poden obtenir de dues formes:

Hi ha una fórmula de correlació lineal de dues variables, la fórmula de **Pearson,** però aquesta no té en compte els rangs dels valors. Per això, la fórmula més emprada en aquest tipus d'estudi és la fórmula **Spearman-Brown,** la qual ens dóna un coeficient de la fiabilitat de partir per la meitat d'una sèrie de valors ordenats pel rang. De tota manera, les estimacions resultants varien de poc.

La fórmula **Spearman:**

$$r_{xx'} = \frac{2r_{hh'}}{1 + r_{hh}}$$

$r_{xx'}$ és el coeficient de fiabilitat

$r_{hh'}$ és la correlació dels resultats ordenats segons el rang obtinguda entre les dues meitats del test.

$1 + r_{hh}$ és la correcció per la longitud.

Dues premisses s'han de conèixer per usar aquest mètode. La primera, com que estem tractant les dues meitats com si fossin tests paral·lels, hem d'assumir que aquestes són **iguals en termes de mitjana i variances** (com a

suposició que podem comprovar). La segona és que hem d'assumir que les dues meitats són experimentalment **independents una de l´altra** (suposició que és més difícil de comprovar).

Una altra aproximació per estimar la fiabilitat partint la prova en meitats és aquella desenvolupada per **Guttman** (1945), la qual no assumeix l'equivalència a les dues meitats, i la qual no requereix computar una correlació entre elles. El coeficient de la fiabilitat de partir per la meitat està basat en la proporció de la suma de les variances de les dues meitats sobre **la variància del test sencer**:

$$r_{xx'} = 2\left(1 - \frac{s^2_{h1} + s^2_{h2}}{s^2_x}\right)$$

on s^2_{h1} i s^2_{h2} són les variances de les dues meitats. Com que aquesta fórmula està basada en la variància del test sencer, ens dóna una estimació directa de la fiabilitat del test sencer. Per tant, malgrat que la correlació entre les meitats era també la base pel coeficient de fiabilitat de Spearman-Brown, l´estimació de partir per la meitat de Guttman no requereix una correcció adicional per a la longitud.

En els estudis de l'estimació de la fiabilitat de partir per la meitat, hi ha moltes maneres diferents segons les quals un test donat podria ser dividit en meitats. Com que no cada partició ens dóna meitats amb característiques exactament iguals en termes d'equivalència i d'independència , els coeficients de fiabilitat obtinguts per diferents particions són susceptibles de variar, per això les nostres estimacions de fiabilitat dependran majorment de la partició particular que utilitzem. Una manera d'evitar aquest problema seria dividir el test en dues meitats de totes les formes possibles i computar la mitjana dels coeficients de fiabilitat basats en aquestes meitats diferents, per obtenir el percentatge

d'aquests coeficients. Aquesta aproximació es torna aviat poc pràctica, ja que el nombre dels coeficients de fiabilitat per a ésser computats s'incrementa de forma exponencial com a una funció del nombre dels ítems en el test. Per exemple, amb 8 ítems hi ha 35 possibles particions i 595 coeficients. Afortunadament, hi ha una forma d'estimar la mitjana de tots els coeficients partits per totes les meitats possibles: les fórmules KR-20 i KR-21, plantejades per l'inspector Rul (2000).

Aquesta aproximació, que va ésser desenvolupada per **Kuder i Richardson** (1937), suposa la computació de totes les mitjanes i de les variances de tots els ítems que constitueixen el test.

La mitjana de l'ítem dicotòmic (que es puntuat com a "correcte" o "incorrecte") és:

- la proporció,simbolitzada amb *p*, dels individus que contesten a l'ítem correctament.
- La proporció dels individus que responen incorrectament a l'ítem és igual a 1-*p*, i es simbolitzatda *q*.
- La variància d'un ítem dicotòmic és el producte d'aquestes dues proporcions, o *pq*.

El coeficient de la fiabilitat donat per la fórmula 20 de Kuder-Richardson (KR-20), es basa en la proporció de la suma de les variances de l'ítem sobre la variància de la puntuació total del test :

$$r_{xx'} = \frac{k}{k-1}\left(1 - \frac{\Sigma pq}{s^2_x}\right)$$

- **k** és el nombre dels ítems en el test,
- **Spq** és la suma de les variances de l'ítem,
- $\mathbf{s^2_x}$ és la variància de la puntuació total del test.

Es requereix que els ítems siguin independents i tinguin un grau de dificultat similar.

La fórmula de **Kuder-Richardson 21** , és com es representa a continuació:

$$\mathbf{r_{xx'}} = \frac{k}{k-1}\left(1 - \frac{M(k-M)}{Ks^2}\right)$$

- $r_{xx'}$ = coeficient de fiabilitat
- k = nombre d' items de la prova
- M = mitjana
- s= desviació típica de les puntuacions de la prova
- s^2= variància

O també es pot operar així:

$$\frac{\mathbf{ks^2 - M(k - M)}}{\mathbf{(k-1)s^2}}$$

Totes dues maneres d'operar ens donen el mateix resultat.

Aquesta fórmula generalment ens donarà un coeficient de fiabilitat menor al que es dona a KR-20. Al igual que el coeficient partit per la meitat de Guttman, les fórmules de Kuder- Richardson estan basades en la variància de la puntuació total, i per tant no requeriran cap correcció per a la longitud.

Tant l'estimació de la partició per la meitat de Guttman com les fórmules de Kuder-Richardson estimen la fiabilitat en base a la proporció de les variances dels components del test –meitats i ítems- sobre la variància de la puntuació total del test.

Cronbach (1951) desenvolupà una fórmula general per estimar la consistència interna que ell anomenà "coeficient alfa", i que sovint s'anomena **"alfa de Cronbach":**

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s^2_i}{s^2_x}\right)$$

- **k** és el nombre dels ítems al test,
- **? s^2_i** és la suma de les variances de les diferents parts del test,
- **s^2_x** és la variància de les puntuacions del test.

En el cas dels ítems dicotòmics, ? s^2_i serà la suma de les variances de l'ítem, ? *pq*, a la fórmula KR-20.

Si les parts són dues meitats, ? s^2_i serà la suma de les dues meitats, $s^2_{h1}+s^2_{h2}$ a la fórmula de Guttman, i l´expressió k/ (k-1) es redueix al valor de 2.

En resumen, la consistència interna d'un conjunt de resultats d'un únic test es pot estimar fent el càlcul de'l coeficient de fiabilitat tenint en compte que:

1. Si utilitzem la partició per la meitat de Spearman-Brown, assumim que les dues parts són equivalents i independents una de l´altra
2. Amb la partició per la meitat de Guttman, solament necessitem assumir que les meitats són independents una de l´altra.
3. En l´aproximació de Spearman-Brown la fiabilitat és *infraestimada* si les meitats no són equivalents.
4. Per ambdues, la fiabilitat és *sobreestimada* si les meitats no són independents una de l'altra.
5. Les estimacions de Kuder- Richardson basades en les variances de l'ítem assumeixen que els ítems en el test són equivalents i independents uns dels

altres. Les fórmules de Kuder-Richardson *infraestimen* la fiabilitat quan els ítems no són equivalents, i la *sobreestimen* quan els ítems no són independents uns dels altres.

6. El error estàndard de la mesura (SEM)

Els coeficients de fiabilitat no ens donen dades sobre la precisió de la mesura per als casos individuals. L'indicador SEM ens ajuda a interpretar els resultats individuals. Es a dir, donat un coeficient de fiabilitat determinat, ¿quin grau de variació es pot calcular en un resultat individual?

Sempre hi haurà una diferència entre el resultat vertader i el resultat obtingut, ja que només es podria esperar que aquestos dos valors coincideixin en el cas que el coeficient de fiabilitat fos 1, cosa gairebé impossible.

$$s^2_e = s^2_x (1- r_{xx'})$$

Fórmula que operada ens dona:

$$s^2_e = s^2_x (1- r_{xx'}) \text{ ®} \quad \boxed{S_e = s_x (1- r_{xx'})^{1/2}}$$

Amb una desviació estàndard de 8 i una fiabilitat de 0.87, per exemple:

$S_e = 8\ (1\text{-}0.87)^{1/2} = 2.88$

Doncs, hi ha una estimació de gairebé 3 punts que pot arribar a haver-hi entre la puntuació obtinguda i la puntuació vertadera.

Per definir el marge o percentatge de confiança de que l'error es trobi 3 punts a dalt o a baix, aquest concepte s'ha de relacionar amb el de la distribució normal.

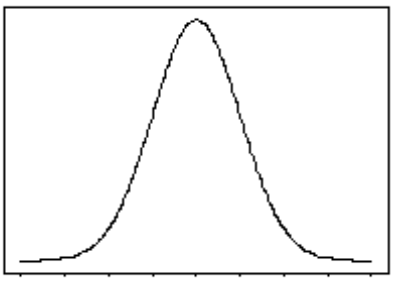
7. La distribució normal

El model de distribució normal és el que idealitza els «histogrames en forma de campana» amb què sovint ens trobem en mesures de distribucions que recullen rendiments en una prova. Aquest model s'assumeix com a el model adient en les proves d'aprenentatge i assoliment d'idiomes (Alderson, 1995, 156).

Quines són les característiques que té aquest model?

Podem pensar que una distribució estadística que recull els valors d'una variable correspon al *model normal* si la idealització del seu histograma (*perfil*) ens mostra una corba simètrica, amb un únic màxim, que coincideix amb la mitjana. Si la corba normal ha d'ajustar una distribució de dades començarem per l'observació del perfil de l'histograma. Hi ha un criteri intuïtiu: la coincidència de la mitjana i la mediana i un altre de numèric: el coeficient d'asimetria (que ha de ser aproximadament 0) per a valorar la simetria d'una distribució de dades empíriques. Aquesta simetria s'ha de veure afectada en les proves d'anglès d'EOI pels percentatges que es requereixen per aprovar (pass mark) que estan al voltant del 60%.

- La forma de la corba normal es caracteritza també per l'existència de dues llargues cues, i per un cert grau d'apuntament que podem valorar amb el coeficient de curtosi:



- Per representar la distribució normal es farà servir aquesta simbologia:

N(μ,s)

μ mitjana s desviació estàndard

Entenem que en una prova en la qual el percentatge per aprovar és d'un 60%, el gràfic del grau d'asimetria s'hauria de desplaçar cap a l'esquerra i el coeficient de curtosi o grau d'apuntament, encara que tingui el mateix valor, surt també esbiaixat i tindrà un perfil més punxegut si el valor de la mitjana és més alt.

- Així doncs, els percentatges de dades que, per a la distribució normal, pertanyen als intervals és:

  $(\mu-\sigma, \mu+\sigma]$, $[\mu-2\sigma, \mu+2\sigma]$, $[\mu-3\sigma, \mu+3\sigma]$,

  - Els esmentats percentatges són ben coneguts i característics d'aquest model, amb aquests valors aproximats per a la distribució:
    68.3% 95.4% 99.7%.

Per intuir si el conjunt de dades empíriques registrat té o no un bon ajust amb el model normal, hem de comparar els percentatges de la distribució dels estadístics amb els que acabem de donar.

Si reprenem l'exemple que hem fet servir per a el càlcul d'error SEM (de 3 punts arrodonits) i afegim que la mitjana de la prova és de 66 **N**(66,8) representa la distribució normal de mitjana 66 i desviació estàndard 8; en aquesta distribució el 68,3% de dades del model teòric estaran entre 58(= - 1s) i 74 (= +1s); el 95.4% entre 50 (= - 2s) i 82 ((= + 2s); finalment pràcticament el 100% entre 42 (= - 3s) i 90 (= +3s).

I si apliquem aquest model a l'anàlisi de l'error, que en l'exemple s'ha estimat en 3 punts (arrodonint el 2.88), podem pensar que **si un individu treu una puntuació de 57 punts:**

un 68.3% de possibilitats que la seva puntuació vertadera estigui entre el 54 i el 60,

un 95.4% de possibilitats que la seva puntuació vertadera estigui entre el 51 i el 63,

un 99.7% de possibilitats que la seva puntuació vertadera estigui entre el 48 i el 66.

Es conclou que **l'error de la mesura és un valor detectable i quantificable** amb uns mínims procediments de càlcul.

OBJECTIUS DEL PROJECTE

1. **Fer un estudi comparatiu dels projectes curriculars** per als ensenyaments/aprenentatges del nivell de primer en els tres centres participants per tal de **definir un constructe** d'acord amb el qual es pugui :
2. **Establir uns descriptors comuns de les activitats comunicatives i les competències concretes**, que ens permeti establir uns criteris de valoració estandarditzats, i uns continguts més o menys fragmentats, amb els quals:
3. Elaborar una prova a partir de **l'anàlisi factorial de validesa**.
4. **Definir les especificacions** de la prova de mesura.
5. Creació del i**nstrument** amb el programa SPSS: introducció i etiquetatge de les variables i assignació de valors
6. Aplicar-la en dues fases:
   - ✓ una d'assaig, sobre un nombre petit d'alumnes per tantejar el funcionament dels items i els indexs de dificultat i discriminació.
   - ✓ una altra definitiva, com a **prova de rendiment comuna sobre una mostra representativa d'alumnes** en centres oficials.
7. **Introduir els resultats obtinguts, com a base de dades** fent servir el programa SPSS i l'instrument ja dissenyat.
8. Anàlitzar amb **eines d'estadística descriptiva** els resultats de la prova i de les dades obtingudes. Tot aplicant, amb caràcter confirmatori, les fórmules de fiabilitat que emmarquen el projecte.
9. Analitzar **la seva funcionalitat com instrument de diagnosi,** les febleses del instrument i les seves possibles causes.
10. Estudi posterior com a factor de **contrast** amb l'avaluació continuada.

RECURSOS EMPRATS

- De la fase de **documentació i anàlisi bibliogràfic**, destacaria l'aportació documental obtinguda principalment en les següents fonts:

1 En llibres d'autors consagrats en l'avaluació de llengües, com ara Alderson, Bachman, Cronbach, Palmer, Ebel,...
Respecte a la bibliografia sobre Metodologia de l'Investigació he consultat les publicacions de Mc Graw desenvolupades per Hernàndez Sampieri & Fernàndez Collado i els textos d'Anàlisi Estadístic de Pardo Merino i César Pérez.

2 Com a recursos virtuals han estat molt útils els articles i resums que estan penjats en internet, com ara la base ERIC, *l'American Educational Research Association, American Psychological Association & National Council on Measurement in Education, Standards of the Educational Testing Service,* etc.

3 També he consultat publicacions de la Biblioteca General del Departament d'Ensenyament (Revista Iberoamericana de Educació, documents de la Comunitat Europea i d'especialistes del Consell Superior d'Avaluació).

- Respecte als **paquets informàtics**, hi ha un ample ventall de programes informàtics que es poden aplicar a l'anàlisi empírica de dades.
Que estiguessin a l'abast, només es van poder fer servir:

1 Dins de la configuració estàndar dels ordinadors, Microsft ofereix el programa EXCEL XP, que originalment va estar un full de càlcul, però, de cada vegada més, incorpora eines d'anàlisi i descripció estadístiques més eficaces.

2 A través del Servei de Tecnologies de l'Informació del Departament d'Ensenyament, el programa MINITAB, que és un programa en català, però amb limitacions per a la creació de bases de dades i per a la manipulació de gràfics. Encara que, sens dubte, ofereix procediments molt potents d'anàlisi.

3 I el programa SPSS, que es fa servir per tractar dades del àrea de les Ciències Socials i de l'Educació a l'Universitat de Barcelona, i que és sens dubte, el que té més abast per processar dades estadístiques.

- **Recursos Humans:**

La col·laboració amb els centres participants va estar fluida al llarg de tot l'estudi. Els professors de set grups classe i els caps adjunts van col·laborar voluntàriament. Em van facilitar la documentació necessària, van aportar correccions i consideracions personals en l'elaboració de l'instrument, van contestar amb cura els qüestionaris previs i posteriors a l'aplicació de la prova, i van permetre que aquesta s'administri dins del seu horari lectiu.

FASES DEL TREBALL

Les fases i la temporització han seguit el plantejament inicial

a. Fase de Planificació, Elaboració i Assaig de la prova:

Planificació

Disseny Del Instrument i Estudi Comparatiu:

Concretament, es va comparar:

1 Els Projectes Curriculars del nivell elemental pel que fa a:

1.1 Els objectius contextualitzats, els coneixements, les competències i les actituds establerts per al nivell de primer EOI.

1.2 Els continguts de fets, conceptes i sistemes conceptuals i procediments i la seva seqüenciació.

2 Les opcions metodològiques.

3 Les pautes, criteris i instruments de l'avaluació.

4 Els llibres de text i gramàtiques, material de suport, documents de tasques i proves previstes o desenvolupades amb anterioritat, qüestionaris o enquestes. Tots els instruments utilitzats per contextualitzar, concretar i coordinar amb coherència l'avaluació del grau de competència que els alumnes del nivell han d'assolir.

Resultats Observats:

Després d'analitzar els PCC pel 1º nivell de les escoles implicades, es va constatat una sèrie de diferències en els diferents apartats.

En principi, l'apartat de **continguts funcionals** L'escola 3 i L'escola 1 tenen el mateix PCC i, amb respecte a l'escola 2, presenten diferències pel que fa a les subdivisions. i.e.:

L'escola 2 distingeix entre Informacions per una banda i Opinions i Valoracions per un altra, mentre que l'escola 3 i l'escola 1 els barregen sota l'epígraf Informacions i Opinions.

D'altres funcions, estan en un dels PCC i no apareix a un altre. i.e.:

al PCC de l'escola 2 no s'inclou les funcions de expressar voluntat o descriure plans pel futur, mentre que en l'escola 3 i l'escola 1 no es té en compte funcions com ara la de donar ordres de manera directa o indirecta o narrar un fet o esdeveniment passat.

Això per esmentar tan sols algunes de les diferències detectades.

De tota manera, els **continguts gramaticals** que expressen aquestes funcions sí que hi són, així com en els llibres de text que es fan servir. Per tant, potser aquestes diferències són degudes a diferents nomenclatures o pressuposicions.

En l'apartat de continguts gramaticals, cadascuna de les tres escoles té un projecte propi i les variacions que he constatat i m'han cridat l'atenció més.

1 Les coses que <u>no hi ha</u> són:

L'escola 1	L'escola 2	L'escola 3
	Especificació dels temps verbals	Posició dels adjectius en el sintagma nominal
	So do I, neither do I, me too.	Count and mass nouns
		Adverbis de freqüència, manera, temps i lloc.Col·locació.

		Sufixos en la formació d'adjectius i adverbis

2 Dels continguts que **hi són en una de les escoles i no apareixen a les altres**:

L'escola 1	L'escola 2	L'escola 3
If sentences- 1º conditonal	Few- little	
Oracions de finalitat amb to	All- both- every- each	
Preposicions com for, with, without	quite	
Pron. Indefinit one		
Verbs amb partícula més freqüents (look at, wait for, listen to,...)		

Moltes de les diferències trobades als textos dels PCCs no són reals, i no van més enllà de diferències de enunciats. En el sentit què després, parlant amb els professors i mirant els llibres de text, els continguts són gairebé els mateixos.

Així mateix, es va afegir un element que no figura en cap PCC però que sí s'ensenya: Conjuncions Adversatives, Coordinatives, Disjuntives, Consecutives i Explicatives.

En qualsevol cas, només es va comptar amb els continguts que, ja sigui de forma verbal o escrita, estaven expressament inclosos en el currículums de totes les escoles.

En l'apartat 2 i 4, no hi havia diferències notables. Totes les escoles tenen un enfocament comunicatiu i unes pautes metodològiques semblants.

Pel que fa a les pautes d'avaluació i als criteris de promoció dels alumnes, l'escola 1 i l'escola 2 fan proves finals d'assoliment, mentre que l'escola 3 aplica proves parcials de caire formatiu i més en la línia de l'avaluació continuada i procesual..

Elaboració

Disseny dels tipus de tasques de la prova

- En principi, la idea era dissenyar una prova que mesuri la competència lingüística dins del àmbit de **l'expressió i la comprensió escrita** que sembla més adequada als propòsits desitjats. Perquè en l'àmbit de l'expressió oral a nivell elemental, el nivell d'interacció es molt baix i el repertori de recursos és massa limitat, les diferències que generen les estratègies comunicatives no ens donen uns resultats tan clarament identificables com ara els resultats obtinguts a través de les proves de mesura en l'àmbit de l'expressió escrita. En aquestes, la competència lingüística, tant la de tipus pragmàtic com la sociolingüística, genera uns resultats més susceptibles de definició dins d'uns paràmetres identificables.
  El constructe es pot acotar amb més claredat. Les funcions, nocions, gramàtica, lèxic i ortografia es poden ajustar a una escala de descriptors/especificacions més fàcil de quantificar i analitzar amb una formulació estadística objectiva.
  En la prova, els tipus de tasca triats són similars als que es fan servir en totes les tres escoles per a les proves d'avaluació formativa o sumativa. Encara que s'han triat i dissenyat per aconseguir el nivell més alt de **fiabilitat i validesa**, és a dir:

- la tipologia de les tasques és **objectiva i quantificable**. Dins del que sigui possible s'ha de dissenyar una prova d'ítems ben estructurats, per tal que el número de respostes possibles per a cadascú es redueix al mínim.

- Per a l'anàlisi de fiabilitat es requereix que els ítems siguin **independents** i tinguin un **grau de dificultat similar**. Tots els ítems han de tenir un valor equivalent. Per això, en l'*Open Dialogue* es va assignar dos ítems per cada apartat: la forma verbal i la "question form". Així, aquest apartat tenía una puntuació de 2 per cada espai buit, però es respecta el criteri de homogeneïtat dels ítems.

- Gairebé totes les facetes de les tasques del test mesuren **les mateixes capacitats**, qüestió fonamental per aconseguir valors acceptables de correlació.
  El *Reading Comprehension* es va plantejar de forma experimental, ja que és un tipus de tasca que no s'aplica per a aquest nivell. Per la qual cosa es va triar un format què tingués un grau de facilitat elevat: col·locar unes frases donades al lloc que els pertoca. Per aquest motiu el format del mètode de la tasca no va assolir l'independència necessària dels ítems i això va quedar molt bé reflectit en l'anàlisi final (és l'única tasca que no assoleix uns coeficients acceptables en l'estudi empíric).

- Amb aquests criteris i, per suposat, amb el del la **representativitat** dels apartats curriculars, es va prendre una decisió ponderada del **número dels ítems**, tot tenit en compte si:
  - ✓ la prova té una **longitud adequada**. Perquè l'efecte de les respostes elegides a l'atzar serà menor i llur fiabilitat serà per tant major que si dissenyem una prova breu.

  - ✓ Malgrat això, tampoc és convenient que la **duració** sigui major al període establert com òptim per activitats que requereixen concentració (aprox 45-50').

- També es important que el **número de tasques i tècniques proposat sigui ample,** per tal d'activar els diferents estils cognitius i destreses. Aleshores, es van dissenyar sis tipus de tasques diferents.

- Els tipus més utilitzats, segons les destreses que es pretenen mesurar, i sempre d'acord amb el tipus de prova habitual en l'ensenyament del nivell elemental d'EOI, són:

1. **Multiple Choice**
2. Matching

3. Information Transfer
4. Ordering Tasks
5. Editing
6. **Missing Word**
7. **Gap Filling**
8. **Error Correction**
9. Ctest
10. Short Answer Question
11. **Question Making**
12. **Reading Comprehension**

En negreta figuren les que es plantegen a la prova (ANNEX 1, Prova Prèvia)

- Es van incloure uns tipus **d'ítems heterogenis però organitzats en blocs homogenis.**

- També hi ha un **petit questionari,** per a la recollida de dades rellevants com ara: l'edat, experiència prèvia, percepció de dificultat, etc.

- **La redacció de les instruccions i dels items** va ser clara i concisa . S'ha de donar unes instruccions clares als examinands sobre com cal respondre, on s'han d'anotar les respostes i les rúbriques han de donar una guia eficaç per les seves respostes en les tasca concreta. A més de les instruccions expressades a la mateixa prova, **les especificacions** que es descriuen posteriorment es van explicar amb claredat durant uns minuts abans de la prova .

- **La elaboració de la clau de respostes correctes q**ue a més de les respostes doni una **puntuació ponderada** amb criteris quantitatius. **L'objectivitat** va estar prioritzada per garantir que estigués per sobra dels judicis discrecionals i la puntuació assignada fos absolutament independent de las valoracions personals.

Dins de l'elaboració de proves, l'apartat referent als **mètodes de les tasques dels ítems** que es triïn és veritablement interessant, ja que hi ha múltiples estudis i orientacions per pautar l'elaboració de tasques.
En aquestes qüestions, es va tenir en compte les precisions de Alderson, Clapham i Walls (1996,45) per tal d'evitar les possibles deficiències, ja que s'ha comprovat que existeix la possibilitat que la puntuació del alumne es vegi afectada segons el tipus de mètode, ja que aquest interactua amb la capacitat en la mesura de l'actuació del examinand (*the method effect)* .
També van ésser fonamentals per al desenvolupament de la prova les nocions i característiques amb què Bachman (1990) va definir les tasques de l'ús de la llengua en el disseny i desenvolupament dels tests. Malgrat el fet que aquest tipus d'estudis estiguin fora de l'abast dels nostres objectius específics, en aquesta recerca s'intenta lligar les tasques de les proves amb les de l'ús de la llengua (TLU), tot seguint les seves opinions sobre les qualitats de les tasques com a factors de l'utilitat de les proves apart de la fiabilitat i la validesa(autenticitat, interacció, impacte i aplicabilitat).

Elaboració d'una escala comuna de descriptors dels continguts i aplicació de la taula d'anàlisi factorial de validesa

(Fitxa Disseny Proves, Rul, 2000):

Actuacions dutes a terme:

- Definir l'àrea curricular a mesurar: es van incloure els blocs de continguts previstos per l'any lectiu sencer. El percentatge d'ítems adreçats a mesurar cada element o bloc de continguts dins de la prova van reflectir la importància assignada a aquest mateix element o bloc de continguts en el PCC.
- Relacionar les metodologies de les tasques dissenyades per mesurar els continguts seleccionats amb les capacitats que implica el seu desenvolupament:
  Aquestes capacitats bàsiques estan relacionades amb els objectius curriculars, i es poden codificar com:

1) **Coneixements de fets, conceptes i terminologia**
2) **Comprensió**
3) **Anàlisi i síntesi**
4) **Aplicació**
5) **Valoració i caracterització**
6) **Organització**

ÀMBIT **1er Curs EOIs**			Tasques
ÀREA CURRICULAR (Global) **Apartats de: gramàtica, morfologia i sintaxi** **(lèxic i ús)**	FITXA DISSENY PROVA	1	Qüestionari
		2	Comprensió Lectora 15 items (no hi ha elements discrets)
		3	Corregir els errors 20 items
	Unitat temps: any 2002-2003	4	Elecció Múltiple 20 items
	Cicle: elemental	5	Omplir els Buits 18 items
		6	Diàleg Obert 20 items

TASQUES	OBJECTIUS (Capacitats) A. oneixements	B. Comprensió	C. Anàlisi i Síntesi	D. Aplicació	E. Valoració i Caracterització	F. Organització	CONTINGUTS CURRICULARS BLOCS O APARTATS DE CONTINGUT		DISTRIBUCIÓ D' ITEMS PER BLOCS O APARTATS DE CONTINGUT Nombre 78		% Proporció	Proporció assignada als PCCs
2	+	+	+		+	+	1.	Article	3	3	3.9	3.2
3	+	+	+	+	+	+	2.	Plural irregular	1	15	19.2	19.4
4	+	+	+		+		3.	Noms contables e incontables	1			
5	+	+	+	+	+	+	4.	Genitiu Saxó	3			
6	+	+	+	+	+	+	5.	Pronoms subjecte-objecte	5			
							6.	Some, any, no- (thing, body,one).	4			6.4
							7.	Sufixos formació noms i adjectius	1			
							8.	Posició-ordre del adjectiu	3	12	15.4	16.5
							9.	Comparatius-Superlatius	3			
							10.	Cardinals, ordinals, intensificadors.	2			
							11.	Much, many, a lot, little,few.	2			
							12.	Possessius- Demostratius	2			
							13.	Partícules Interrogatives	7	7	9	3.2
							14.	Adverbis de manera, freqüencia, temps i lloc	3	3	3.9	3.2
							15.	Preposicions de temps	2	6	7.7	9.7
							16.	Preposicions + verb més habituals	2			
							17.	Preposicions de lloc i direcció.	2			
							18.	Verbs lèxics	2	27	34.6	38.7
							19.	Verbs auxiliars	4			
							20.	Present simple	3			
							21.	Present Progressive	1			
							22.	Past Simple (regular/irregular)	6			
							23.	Future- present progressive	1			
							24.	Future- going to	4			
							25.	Future- will	1			
							26.	Imperative	1			
							27.	Can- must- have to	1			
							28.	Love- like (would like)- hate- prefer	3			
							29.	Gerundi- Infinitiu	2	2	2.6	3.2
							30.	Conjuncions Adversatives, Coordinatives, Disjuntives, Consecutives i Explicatives	2	2	2.6	?

Elaboració de les Especificacions de la Prova

Les especificacions de la prova són un document fonamental: el marc descriptiu on s'estableix la definició bàsica per al coneixement de totes les persones involucrades.

Quins elements definitoris es poden incloure-hi?

Totes les explicacions que ajudin a fer-se una idea clara i precisa de la prova: sobre l'instrument i el seu propòsit, els agents implicats, els detalls del format, l'administració, etc. Com a mínim, s'hauria de definir:

1. El objectiu de la prova
2. Els continguts
3. Les habilitas que es mesuren
4. Participants
5. Els destinataris
6. Les seccions
7. Els tipus de tasques
8. Els mètodes de les tasques
9. Criteris de puntuació
10. Duració
11. Feedback

A continuació, veurem aquests principis plasmats en les especificacions que es van enviar als professors i es van explicar als alumnes que van participar-hi.

Especificacions Prova de Rendiment

"Validació de Proves de Rendiment d'Anglès en el Nivell Elemental" és un projecte d'estudis sobre la fiabilitat y la validesa de proves en tres EOIs de Barcelona[1]. És decisiva la col·laboració dels professors i dels caps de nivell de primer, als quals voldria agrair pel seu interès i participació activa.

Finalitat: aquesta prova està dissenyada com a instrument de recerca. És, d'una banda, una eina de contrast de l'avaluació continuada. De l'altra té com a objectiu la diagnosi dels encerts i febleses d'alguns dels instruments de mesura dels coneixements que es fan servir a les escoles d'idiomes de Barcelona, capital.

La prova s'administrarà a tres escoles i ens donarà uns resultats a partir dels quals es podran aplicar dos tipus d'anàlisi:

1) De **fiabilitat**: és a dir, fins a quin punt la prova és consistent i adequada per mesurar els coneixements dels alumnes i no ens dóna una informació esbiaixada o d'errors que siguin deguts a la manca de consistència de la pròpia prova. Aquesta anàlisi es fa amb criteris estadístics i amb els programes informàtics Minitab i SPSS.
   Amb aquesta mostra es pot definir una sèrie de paràmetres com ara la desviació estàndard; el perfil de la mediana, mitjana i la moda; l'índex de dificultat; els coeficients de correlació, la distribució; etc. Així com una sèrie de gràfics que ens permetrà copsar de forma intuïtiva i clara l'assoliment dels coneixements i les febleses del aprenentatge.

[1] **que es pot dur a terme gràcies a una llicència concedida pel Departament d'Ensenyament de la Generalitat de Catalunya**

L'objectiu més important és treure els coeficients de fiabilitat i les correlacions, que ens permetran confirmar la fiabilitat o no de l'instrument de medició.

2) De **validesa**: és a dir, fins a quin punt les interpretacions que es fan dels resultats obtinguts són adients i significatives. La validació és la demostració empírica què els resultats de la prova reflecteixen l'habilitat lingüística a mesurar i no cap altra cosa. Aquesta demostració es fa tot seguint una anàlisi factorial proposat per l'inspector i director d'aquesta recerca, Jesús Rul,, i s'aplica en la fase de disseny i, amb posterioritat a la prova, s'ha de constatar.

Amb l'anàlisi factorial previ es comprova l'equilibri entre el continguts curriculars (de fets, conceptes y sistemes conceptuals/ procediments) i els de la prova. També es verifica la proporció de les capacitats (cognoscitives, aplicatives i valoratives) que implica la prova i l'assignació de punts.
La prova prèvia ens ha permès introduir canvis i correccions en els ítems de la prova final: els que no havien obtingut uns paràmetres acceptables o les rúbriques
i pautes per a l'administració que no funcionin s'han modificat i també s'ha fet un ajust de la puntuació.

Nivell: elemental.

Els llibres de text que s'utilitzen són:
New Headway Beginner & Elementary (OUP), *English File* (OUP) i *Milestones* (OUP).

Característiques de la prova: És una prova escrita, de mesura d'estructures gramaticals, comprensió lectora i expressió escrita molt guiada. Com que l'objectiu de la prova és l'anàlisi empíric, el tipus de resposta ha d'estar molt ben definit i els criteris de correcció seran objectius. Un percentatge alt només té una única resposta correcta i no pas uns criteris

més o menys flexibles o subjectius de correcció. Malgrat això, l'èmfasi dels continguts triats s'ha d'adreçar a l'ús comunicatiu de la llengua.

Constructe: el constructe ve definit pels PCCs de les tres escoles que hi participen en el projecte: els continguts que hi ha en els apartats de pragmàtica i de gramàtica, morfologia, sintaxi i lèxic. Els elements que no són a cap de les tres programacions no s'hi han inclòs.

Destinataris: Alumnes que hagin cursat el primer curs, que hi participin de forma voluntària. El nombre mínim d'alumnes per la mostra és de 200.La prova prèvia es durà a terme amb els cursos intensius, al gener de 2003 i la prova final tindrà lloc al maig.

Repercussió afectiva sobre els alumnes: com que , encara que sigui una prova d'assoliment, els resultats no seran referència per "aprovar" o "suspendre" ningú, és previsible que la prova no els suposi cap mena de tensió. Més aviat, espero que tinguin interès a participar-hi, ja que obtindran unes dades "paral·leles" de contrast amb l'avaluació oficial i una informació detallada de com funcionen els instruments de mesura.

Seccions:

1) Qüestionari sobre les característiques del alumne y els detalls rellevants per l'aprenentatge.
2) Comprensió lectora.
3) Correcció d'errors.
4) Diàleg obert
5) Omplir espais buits
6) Elecció múltiple

Tasques:

➢En la **Comprensió lectora** els alumnes han de llegir un text d'entre 200 i 250 paraules.La tasca consisteix a col·locar cinc frases tretes del text al lloc que els pertoca. En la versió 2 la tasca serà decidir si deu frases relacionades amb la lectura són correctes o incorrectes.

➢En la **Correcció d'Errors i Afegir una Paraula**, tant a la versió 1 com a la 2, els alumnes hauran de trobar-hi una errada o afegir-hi una paraula.

➢En el **Diàleg Obert**, en les dues versions, la tasca consisteix a omplir una conversa entre dues persones. S'han d'elaborar les preguntes adients a les respostes que ja vénen donades.

➢En l'exercici d'**Omplir Els Espais Buits**, s'ha d'escriure només una paraula en cadascú dels deu espais buits en un text de 150 paraules. Hi ha dues versions.

➢La tasca d'**Elecció Múltiple** (V1 i V2) són 20 frases que s'han de completar triant una de les tres opcions que s'ofereixen.

Rúbriques: Les instruccions porten un dibuix: 🕮 i es donen en català.

Administració: La prova té una durada aproximada de una hora com a màxim, sense descans. Les explicacions prèvies seran, com màxim, de deu minuts, i sempre en la llengua oficial o en castellà. El lloc serà les mateixes aules de classe, i com que hi ha dues versions no cal un espai més gran.

Puntuació:

SCORE → / TASCA ↓	INPUT	OUTPUT	nº items	PUNTUACIÓ PER ÍTEM	TOTAL 70
COMPRENSIÓ LECTORA	Text en anglès de 200-250 paraules	col·locar 5 frases al lloc que els pertoca	5	2	10
CORRECCIÓ D'ERRORS/MISSING WORD V1 I V2	10 frases en anglès	Corregir 5 frases i afegir una paraula en altres 5	10	1	10
DIÀLEG OBERT V1 I V2	Diàleg en anglès de 20 línies	Redactar 10 preguntes en anglès	10	2	20
OMPLIR ESPAIS BUITS V1 I V2	Text en anglès de 150 paraules	Omplir 10 espais buits amb una sola paraula en anglès	10	1	10
ELECCIÓ MÚLTIPLE V1 I V2	20 frases en anglès	Completar-les amb una de les tres opcions	20	1	20

Feedback: Els alumnes tindran una sessió informativa per donar-los els resultats i l'anàlisi estadístic de les proves.

Creació Del Instrument Spss

Amb les dades anteriors vam encetar la creació del i**nstrument** amb el programa SPSS: introducció i etiquetatge de les variables i assignació de valors.
Aquesta fase va estar molt feixuga al començament, per l'esforç formatiu i de familiarització amb el software. A més a més, es requereix que el desenvolupador de la prova tingui clar els continguts, el tipus i el mètode de tasques, l'organització per blocs, l'assignació de punts, etc. Que és el que es va delimitar als apartats anteriors.
Una vegada fet l'instrument és molt fàcil de traslladar a format Minitab o Excel, que ens ofereixen d'altres eines, potser menys potents des del punt de vista estadístic però molt útils pel que fa a visualització de gràfics(per exemple el de tija i fulles .mtw)o l'actualització de dades que facilita Excel amb l'arrosegament de fórmules i les taules actives.

L'administració de les proves com assaig (pretesting)

L'aplicació experimental es va fer en els grups intensius de les escoles 1 i 2 la segona setmana de gener.
Això és un requisit imprescindible, ja que cal fer una aplicació experimental per detectar qualsevol deficiència en el disseny o en les fórmules de càlcul (grau de discriminació, grau de dificultat, assignació de valors,...) També es van introduir modificacions i millores en base a judicis obtinguts amb pautes d'observació, com ara les entrevistes amb els professors implicats.
Donat el nombre reduït d'alumnes als què es va poder passa la prova va estar impossible calcular els coeficients de fiabilitat perquè no és una mostra representativa.

Introducció de dades en la base de dades SPSS i Minitab:

Una vegada corregides les proves(ANNEX 2, notes pretest)hem de endegar la creació de la nostra base de dades. L'introducció de dades és un procés llarg i requereix rigor, però és una eina de treball fonamental de l'anàlisi empírica.

Encara que el nombre de dades (27 alumnes) de la prova prèvia no era rellevant des de el punt de vista estadístic, va estar molt útil per a familiaritzar-se amb la creació i l'ús dels programes SPSS, Excel i Minitab. (ANNEX 3, base dades mtw i spss).

Revisió de les dades obtingudes:

Una vegada duta a terme l'aplicació experimental de la prova es va tenir ja bona informació per avaluar la validesa de les especificacions i si s'havien assolit els seus objectius, en el sentit de si havia aconseguit que quedi reflectit amb precisió el grau de competència comunicativa dels alumnes; si el número la tipologia i la redacció dels ítems eren adequats; si les instruccions i rúbriques eren prou clares; si els recursos emprats eren adients; si les escales de puntuacions i els criteris de qualificacions havien funcionat de la forma prevista; si les mesures per l'administració eren correctes,...

Descripció Gràfica Global de les dades:

Curiosament, se'ns presenta una distribució bimodal, més clarament amb la versió primera o v1.
Això ens revela l'existència de dos grups amb un grau d'habilitat diferent. I en el registre de puntuacions es confirma l'observació.
Davant una distribució bimodal, per tant cal aprofundir en l'estudi dels paràmetres de centralització o de tendència central, que intenten explicar a través d'un sol valor quina és "la tendència majoritària" dels valors observats en la col·lecció de dades que s'analitza. La **mitjana** ja es va comentar al Marc Teòric.

Respecte a la **moda** i la **mediana**:

1. **La moda (mode)**

La moda es defineix com el valor de la distribució que ha estat observat amb una freqüència més elevada. Aquest paràmetre és molt intuïtiu però no acostuma a

tenir transcendència estadística. De tota manera és un indicador a tenir en compte quan es produeixen distribucions de valors anòmales.

2. La mediana (median)

La mediana es defineix com aquell valor tal que, si s'ordenen els valors de la distribució, ocupa el lloc central en aquesta ordenació. És el valor de la distribució per al qual la freqüència relativa acumulada és del 50%

3. Estudi comparatiu de la mitjana i la mediana

- la mediana presenta limitacions pel fet de no tenir en compte totes les dades de la distribució (el canvi d'una dada no té perquè fer canviar el valor d'aquest paràmetre).

- precisament pel que acabem de comentar la mediana és un paràmetre rellevant en cas que la distribució presenti dades singulars, justament el contrari de la mitjana, que es veu fortament influïda si ens trobem amb dades singulars o valors atípics.

- una manera intuïtiva de "mesurar" el grau de simetria d'una variable numèrica és la de comparar els valors de la mediana i la mitjana. Efectivament, si la distribució és totalment simètrica la mediana i la mitjana coincidiran i, en canvi, la distribució diferirà tant més d'un model simètric quant més distanciades estiguin la mediana i la mitjana, de tal manera que "la cua més allargada" es presentarà cap el cantó de la distribució on es trobi la mitjana.

V1

Statistics

SUMA TOTALDE PUNTOS

N	Valid	14
	Missing	0
Mean		33,43
Median		35,50
Mode		47

V2

Statistics

SUMA TOTAL

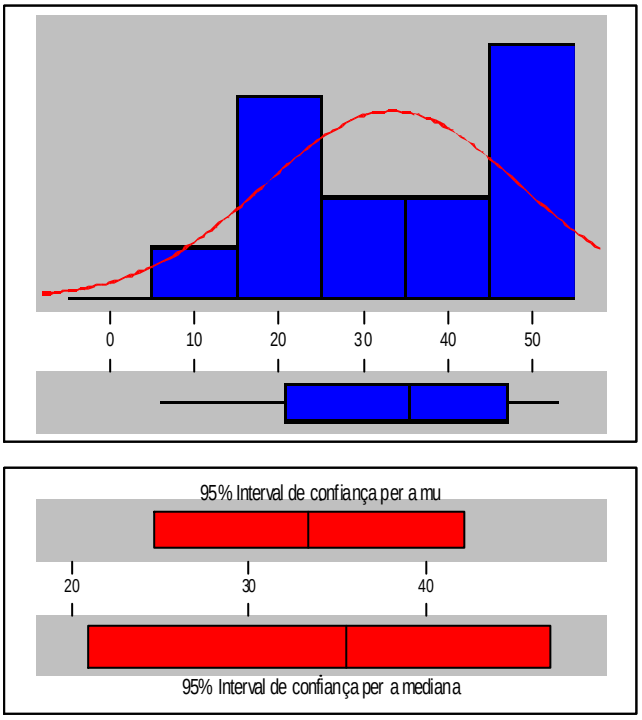
N	Valid	13
	Missing	0
Mean		35,77
Median		38,00
Mode		26

Veiem que la mitjana i la mediana presenten valors semblants, o sigui que hi ha una “certa” aproximació al model normal.. Si tenim en compte que el total possible eren 60 punts, doncs es pot concloure que són valors acceptables (encara que molt millorables).

Ens crida l’atenció la moda que en tots dos casos s’allunya del altres dos valors, però aquest no és un valor estadísticament rellevant, com tampoc és una mostra representativa. Només reflecteix un cert perfil bimodal, es a dir que hem passat la prova a dos grups amb un nivell d’habilitat diferenciat. Això és obvi si consultem els resultats o mirem els gràfics.

Amb el programa Minitab es poden demanar gràfics i anàlisi descriptiu dels resultats, que ens permetran visualitzar d'un cop la distribució, comprovar si la desviació estàndard i la mitjana tenen valors acceptables.

Versió 1

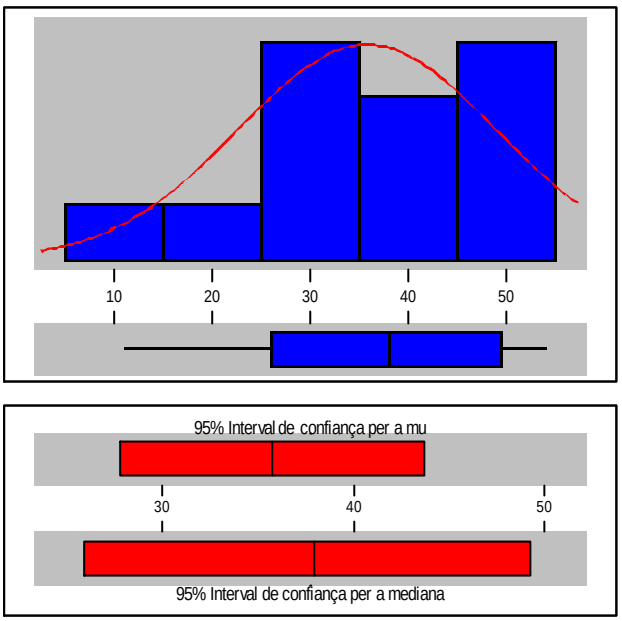


Variable: C59

Prova de normalitat d'Anderson-Darling	
A-quadrada:	0,548
valor-p:	0,129
Mitjana	33,429
Desv. estd.	15,119
Variància	228,571
Asimetria	-0,255
Curtosi	-1,516
n de dades	14,000
Mínim	6,000
1er. quartil	20,750
Mediana	35,500
3er. quartil	47,000
Màxim	53,000
95% Interval de confiança per a mu	
24,699	42,158
95% Interval de confiança per a sigma	
10,960	24,357
95% Interval de confiança per a mediana	
20,949	47,000

Versió 2

Anàlisi estadística



Variable: C59

Prova de normalitat d'Anderson-Darling	
A-quadrada:	0,383
valor-p:	0,343
Mitjana	35,769
Desv. estd.	13,154
Variància	173,026
Asimetria	-0,193
Curtosi	-1,311
n de dades	13,000
Mínim	11,000
1er. quartil	26,000
Mediana	38,000
3er. quartil	49,500
Màxim	54,000
95% Interval de confiança per a mu	
27,820	43,718
95% Interval de confiança per a sigma	
9,432	21,714
95% Interval de confiança per a mediana	
26,000	49,315

La dificultat de la prova:

La dificultat de la prova ha de ser mitjana, és a dir que l'han de poder superar més de la meitat dels que la contesten.

Es considera un FV bo quan el contesten encertadament al voltant del 60% dels alumnes.

Es considera que una prova és fàcil si la superen el 90% dels alumnes i es considera difícil si tan sols ho fan el 10%.

Hi ha diversos procediments de càlcul del índex dificultat. En aquest cas es va elaborar una taula de freqüència SPSS (ANNEX 4, taula freqüències) i es va aplicar el següent criteri contrastat estadísticament:

els ítems amb un índex de facilitat (FV)de entre 0.7 i 0.3 són els que poden arribar a uns coeficients més alts de discriminació (DI).

Per la qual cosa el 80% dels ítems presenten un grau de dificultat dins d'aquesta franja. Un 5% té un grau inferior i un 15% un grau més alt (Rul, 2000).

Grau de discriminació dels ítems:

Un ítem ha de discriminar entre els examinands bons i els dolents. El **coeficient de discriminació** dels ítems està estretament lligat amb la desviació típica. Hem de vigilar aquest paràmetre perquè, per les aplicacions fetes abans d'endegar aquest projecte he pogut apreciar que l'índex de desviació típica es fonamental a l'hora d'obtenir un grau de fiabilitat acceptable.

Si s'observa, la fórmula de càlcul del coeficient de fiabilitat té un denominador que és la xifra de la desviació típica elevada al quadrat. Per tant, d'un denominador que sigui 1.2^2, obtindrem 1.4; mentre que d'un 8^2 ens sortirà un denominador de 64, no cal esforçar-nos gaire per preveure les variacions dels resultats que podem arribar a obtenir. Així doncs, les conseqüències estadístiques d'una desviació típica baixa i d'un índex de discriminació baix (DI) poden ser catastròfiques, des d'un punt de vista estadístic i des d'un punt de vista pedagògic

Fent una comparació manual de frequències amb Minitab entre els 10 alumnes amb la millor nota i els deu alumnes amb la nota més baixa, es van mantenir el ítems que havien estat més discriminatius, es van treure els que van resultar massa discriminatius. L'índex màxim de discriminació és 1, i es considera acceptable un valor que com mínim sigui de 0.5.

Per exemple: si un ítem ha estat contestat bé per 8 alumnes del grup superior i per 2 alumnes del grup inferior, doncs el càlcul serà 0.8 – 0.2= 0.6 DI, que ja és un valor acceptable.

Registre de Canvis

1. **Organització:**

Dates:

1. Fixar les dates amb dos mesos d'antelació i per escrit.
2. Amb un mes d'antelació donar el esborrany de la prova definitiva per introduir possibles canvis.
3. Visita explicativa prèvia a l'aula.
4. Fixar amb els caps de nivell la data per al lliurament de notes, potser prefereixen que es donin abans de l'avaluació continuada.

Administració:

1. Dins l'horari de classe.
2. Amb 60 minuts n'hi ha prou. De fet, la gran majoria dels alumnes acaba abans de l'hora.

Lliurament de les notes:

En una sessió conjunta per escola, en la sala d'actes per exemple, amb totes les classes que hi participin.

2. Continguts:

Qüestionari:

Canviar: Quant temps? per DURANT QUANT TEMPS? i afegir-hi FA QUANT DE TEMPS?

Instrument d'anàlisi estadístic:

Per tal de que l'anàlisi estadística sigui més fàcil de tractar i dongui uns resultats acurats, s'hauria de:

1. Canviar la numeració per blocs dels ítems per una successiva, del 1 al 60.
2. La puntuació assignada ha de referir-se a números enters, ja que el programa no estableix referències creuades amb els mig punts o decimals. I a més la puntuació ha de ser més o menys equivalent. Redefinició de l'*Score.* (ANNEX 5, Score Definitiu)
3. La població de la mostra hauria d'anar al voltant dels 200 alumnes.

Aplicant la **fórmula de la n i de la e** amb l'Excel tenim que:

- Si considerem els alumnes de 1er d'EOI com a una població finita, és a dir només els 900 que hi són a les tres escoles, tenim un coeficient de 0.94 de seguretat o **nivell de confiança** de les nostres dades amb una mostra de 206 persones.

➢ Si considerem que són tots els alumnes de nivell elemental de totes les EOI, doncs la població és infinita i el coeficient d'error possible de la nostra mostra pujaria a un 7%, que trobo força raonable.

Simulació per a poblacions finites

Z=	1,960
p=	0,500
q=	0,500
N=	900,000
e=	0,060
n=?	**205,96**

N=	900
n=	203
e=?	**0,06**

Simulació per a poblacions infinites

Z=	1,96
n=	203
e=?	0,07

- **Tipus d'exercici:**

  Un professor hem va suggerir de introduir un exercici de "*Answer The Following Questions*" i reduir el de "*Make questions to the following answers".* Però com que l'exercici de "*Make questions to the following answers*" va donar els valors més alts en l'anàlisi de correlacions amb el resultat global, és a dir que donava un alt grau de fiabilitat (ANNEX 6, Correlacions OD)
  El vaig deixar només amb uns lleugers canvis.
  La comprensió lectora va resultar molt fàcil, però es va deixar tenint en compte que hi ha el criteri que recomana que un 15% dels ítems tinguin un grau baix de dificultat.
  L'inspector Rul hem va recomanar que inclogués tots els continguts funcionals (ANNEX 7, llista de funcions comunes) Per la qual cosa vaig haver de fer uns petits canvis en l'apartat de selecció múltiple que van acabar d'arrodonir l'instrument.

- **Ítems:**

  Canviar els items de:

DIFICULTAT ALTA	**DIFICULTAT BAIXA**	**DISCRIMINACIÓ BAIXA**
Question tag 1.5	The first 2.5	Will/Going to 1.5 i 2.5
Children are at school 2.1	Dont´t shout 2.5	Our 2.2
Expensive 2.1	Are crossing 1.5	Didn't 2.2
	Eighty-two 1.5	
	Some 1.5	
	She 2.2	
	Stayed 2.2	
	In 1.2	

3. Persones Involucrades:

Els caps d'estudis, per a facilitar-me la disponibilitat de la sala d'actes per al lliurament de notes. Els caps de nivell, per negociar els detalls de organització, administració, tipus d'exercici i continguts curriculars.

4. Anàlisi:

Estudi Comparatiu: Vigilar la privacidad de les dades. Evitar l'establiment de rànkings.

B.Fase de Reformulació i Elaboració del Document Final

Amb tots els paràmetres abans esmentats, es va reformular, reinsertar els canvis en l'instrument estadístic i reeditar el document de la prova inicial, incloent-hi totes les correccions oportunes. (ANNEX 8, Prova definitiva)

Aplicació de la Prova. Correcció i Introducció de les Dades

Aquesta fase va estar molt similar a la fase prèvia, amb una diferència: que aquesta vegada es va aplicar a una mostra representativa: 203 alumnes. Això va significar molta més feina pel que fa a la correcció (ANNEX 9, Notes) i l'introducció de dades. (ANNEX 10, base dades v1 i v2 definitiva)
Però la feina va valer la pena, perquè va suposar un objecte d'anàlisi riquísim. De fet va estar difícil seleccionar els tipus d'anàlisi més adients, ja que les eines que tenim en els paquets estadístics ens donan una quantitat d'informació inabastable.
Bàsicament, he triat els anàlisis que confirmen la validesa i la fiabilitat de la prova i també aquells que fan una descripció fidel dels resultats obtinguts, posant en relleu alguns detalls que he trobat interessants per a utilitzar com a dades objectives per a una avaluació del estat de l'ensenyament al nivell d'anglès elemental a les EOI.

C. Fase Avaluació de Dades de la Prova de Rendiment

Anàlisi de Fiabilitat i Validesa:

Coeficients de Fiabilitat:

V1

NOMBRE DE CASOS = 97

EQUAL-LENGTH SPEARMAN-BROWN = 0,9515

GUTTMAN SPLIT-HALF = 0 ,8056

ALPHA CRONBACH= 0,9204

V2

NOMBRE DE CASOS = 106

EQUAL-LENGTH SPEARMAN-BROWN = 0,8109

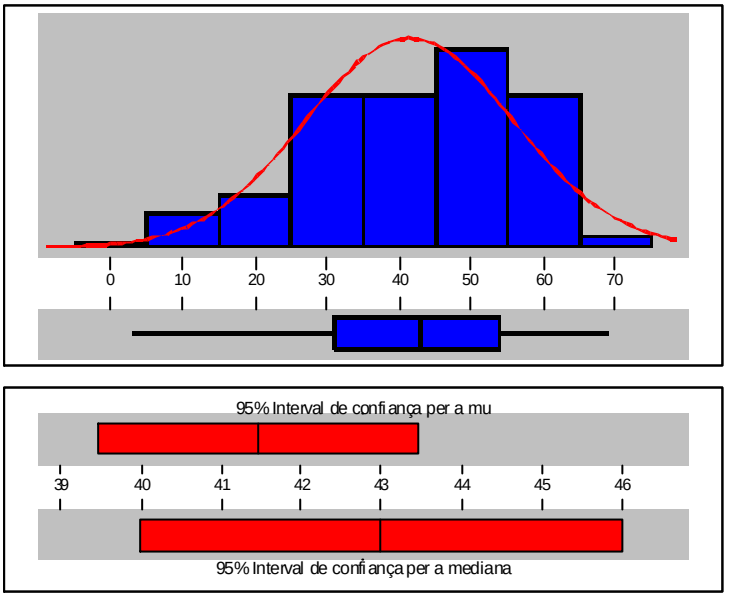
GUTTMAN SPLIT-HALF = 0 ,7766

ALPHA'S CRONBACH= 0,8912

Vaig aplicar la fórmula KUDER RICHARSON 21 al total dels resultats:

$$\frac{ks^2 - M(k - M)}{(k - 1)s^2}$$

Anàlisi estadística



Variable: C1

Prova de normalitat d'Anderson-Darling	
A-quadrada:	1,303
valor-p:	0,002
Mitjana	41,458
Desv. estd.	14,390
Variància	207,061
Asimetria	-0,370
Curtosi	-0,593
n de dades	203,000
Mínim	3,000
1er. quartil	31,000
Mediana	43,000
3er. quartil	54,000
Màxim	69,000
95% Interval de confiança per a mu	
39,467	43,450
95% Interval de confiança per a sigma	
13,113	15,944
95% Interval de confiança per a mediana	
40,000	46,000

Total dels resultats 1

$$\frac{70.207{,}061 - 41{,}458\ (70 - 41{,}458)}{(70 - 1)\ 207{,}061}$$

El resultat corregit amb dos llocs decimals és: 0,932

L'Error de Mesura:

$$S_e = s_x\ (\ 1 - r_{xx'})^{1/2}$$

També l'aplicarem sobre la puntuació total:

$$S_e = 14{,}390\ (1 - 0{,}932)^{1/2}$$

= 3,75 (5,36% del total)

O sigui que tenim una franja de 3.75 punts abaix o adalt que hem de respectar com a error de la nostra mesura.

És interessant que el marge d'error comprovat coincideix amb el marge d'error plantejat com a hipótesis amb **la fórmula de n i de e de la mostra.**

Taules de Freqüencies:

(ANNEX 11, taula freqüències)

Les quals confirmen que no hi ha hagut cap ítem amb un coeficient inferior al 30%. Això vol dir que tots els ítems han tingut un grau de dificultat acceptable.

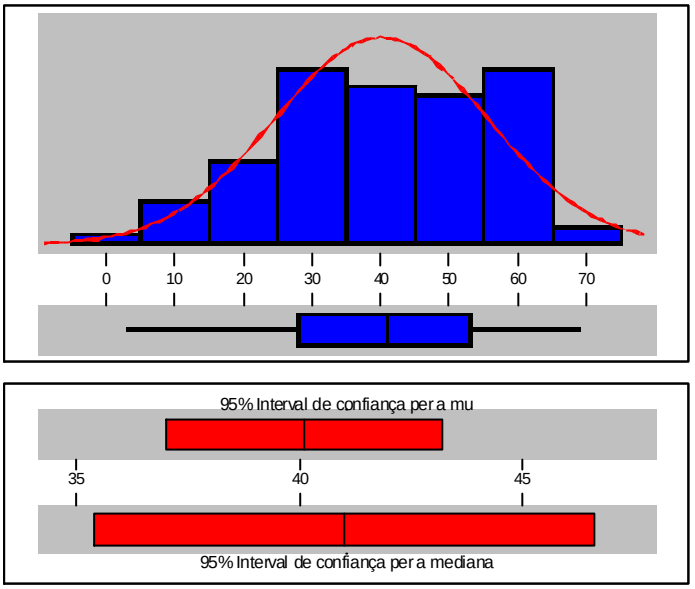
Les Correlacions:

(ANNEX 12, Correlacions)

He tret correlacions **Pearson i Spearman,** encara que vaig subratllar les segones, que són les més adients per aquest tipus d'estudi. Ho he fet grup per grup, per demostrar que hi havia uns coeficients molt alts en tots els grups, independentement del fet que els resultats haguessin estat millors o pitjors. L'única prova que ha obtingut correlacions per sota del nivell desitjable ha estat la Comprensió Lectora, malgrat el fet que va tenir un FV alt. Això s'ha degut a la manca d'independència dels ítems.

Gràfics Descriptius:
V1

Anàlisi estadística

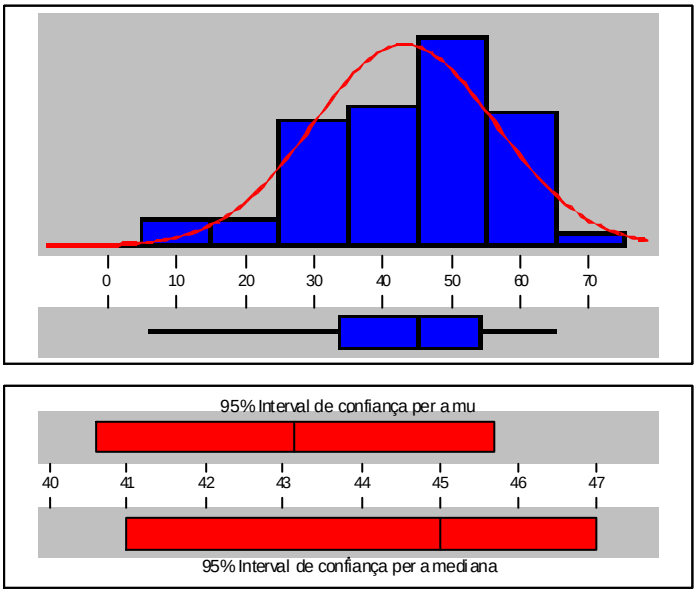


Variable: C1

Prova de normalitat d'Anderson-Darling	
A-quadrada:	0,684
valor-p:	0,072
Mitjana	40,103
Desv. estd.	15,511
Variància	240,593
Asimetria	-0,184
Curtosi	-0,918
n de dades	97,000
Mínim	3,000
1er. quartil	28,000
Mediana	41,000
3er. quartil	53,000
Màxim	69,000
95% Interval de confiança per a mu	
36,977	43,229
95% Interval de confiança per a sigma	
13,593	18,064
95% Interval de confiança per a mediana	
35,395	46,605

V2

Anàlisi estadística



Variable: C1

Prova de normalitat d'Anderson-Darling	
A-quadrada:	0,657
valor-p:	0,084
Mitjana	43,132
Desv. estd.	13,228
Variància	174,973
Asimetria	-0,538
Curtosi	-0,105
n de dades	106,000
Mínim	6,000
1er. quartil	33,750
Mediana	45,000
3er. quartil	54,000
Màxim	65,000
95% Interval de confiança per a mu	
40,585	45,680
95% Interval de confiança per a sigma	
11,655	15,295
95% Interval de confiança per a mediana	
41,000	47,000

Les dues distribucions s'ajusten a una model teòric de distribució normal, amb un petit grau d'asimetria negatiu perquè els valors registren un esbiaix cap a l'esquerra i en la primera versió la curtosi o el grau d'apuntament és negatiu, donant-nos un perfil platocúrtic o amb menys punxa del model.

La coincidència mitjana/ mediana també són dades que ens confirmen l'ajust amb el model normal.

Els valors de la mitjana està al voltant del 60% del total, que és el valor fixat com a nota d'aprovat estàndard. Per la literatura existent i com a mitjana del criteri dels professors.

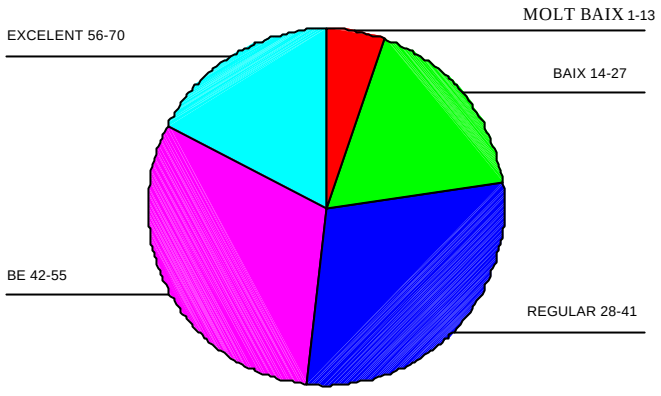
En general la bondat de l'ajust i la distribució de les dades de la versió 2 és més satisfactòria. És interessant que el 50% dels valors registrats, que es troben des del 1er al 3er quartil, estan entre els 34 i els 54 punts.

Distribució de les puntuacions:

Les bandes per a les notes es van establir a partir de la mitjana de les desviacions estàndar de les dos versions, 14 punts arrodonits.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Molt baix 1-13	5	5,2	5,2	5,2
	Baix 14-27	17	17,5	17,5	22,7
	Regular28-41	28	28,9	28,9	51,5
	Bé 42-55	30	30,9	30,9	82,5
	Excel.lent 56-70	17	17,5	17,5	100,0
	Total	97	100,0	100,0	

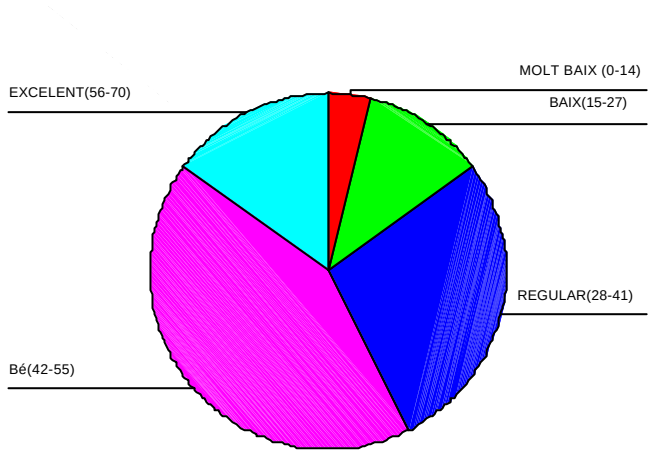
Notes v1



V2

Notes

Notes V2

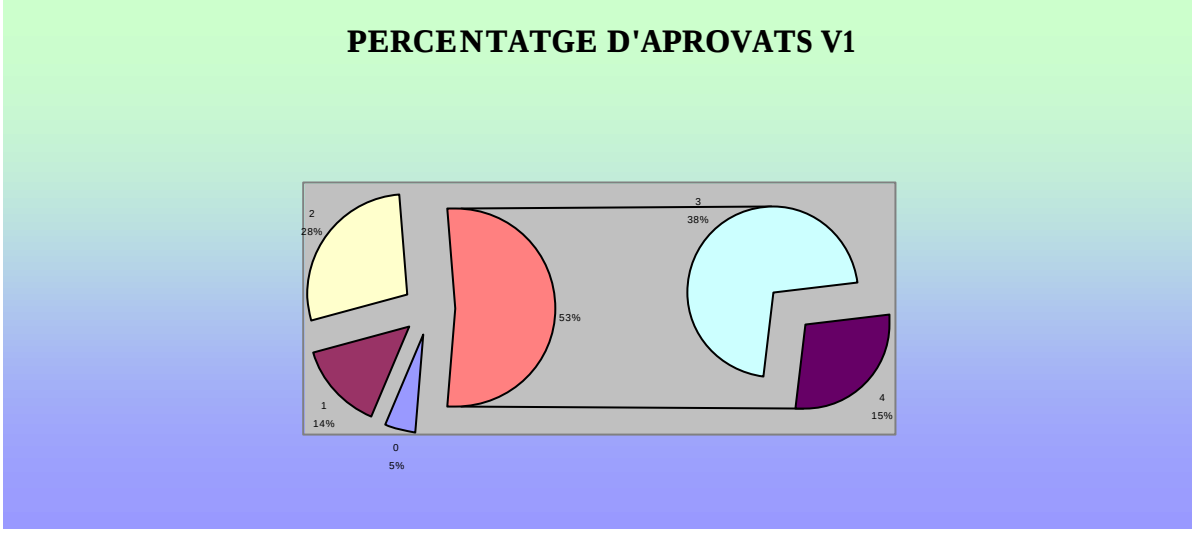


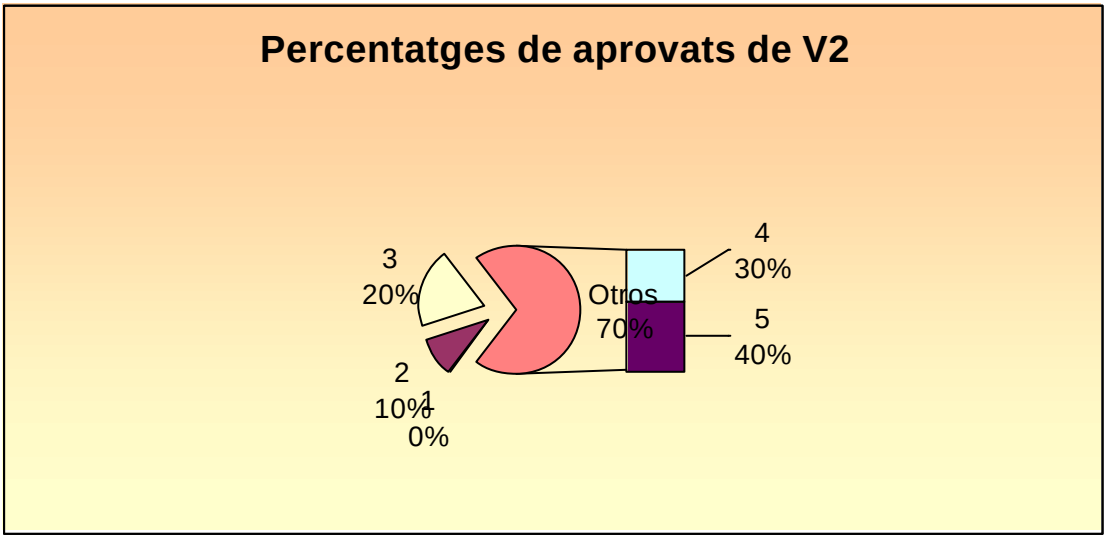
			Cumulative Percent
Vali		ɜnt	
		3,8	3,8
		1,3	15,1
		7,4	42,5
		2,5	84,9
		5,1	100,0
		0,0	

Percentatges per a cada valor

Percentatges d'aprovats

Els percentatges d'aprovats són més que acceptables:





Taules Creuades

Amb el procediment de taules creuades es poden obtenir infinitat de detalls. Relacionant els tipus d'exercici, amb el tipus d'alumnat (experiència prèvia, sexe, edat, etc.). Com ara el fet que l'any de naixement influeix molt en la nota obtinguda:

	ANY NEIX	Molt baix-	Baix	Regular	Bé	Excel.lent	Total
	41-50		2	1	1		4
			11,8%	3,6%	3,3%		4,1%
	51-60	3	4	3	1	2	13
		60,0%	23,5%	10,7%	3,3%	11,8%	13,4%
	61-70	1	6	7	10	2	26
		20,0%	35,3%	25,0%	33,3%	11,8%	26,8%
	71-80	1	4	11	14	11	41
		20,0%	23,5%	39,3%	46,7%	64,7%	42,3%
	81-90		1	6	4	2	13
			5,9%	21,4%	13,3%	11,8%	13,4%
Total		5	17	28	30	17	97
		100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Versió 2

	ANY NEIX	Molt baix-	Baix	Regular	Bé	Excel.lent	Total
	41-50			1	1		2
				3,4%	2,2%		1,9%
	51-60		4	7	5	1	17
			33,3%	24,1%	11,1%	6,3%	16,0%
	61-70	3	3	10	15	3	34
		75,0%	25,0%	34,5%	33,3%	18,8%	32,1%
	71-80	1	3	9	16	10	39
	71-80	1	3	9	16	10	39
		25,0%	25,0%	31,0%	35,6%	62,5%	36,8%
		25,0%	25,0%	31,0%	35,6%	62,5%	36,8%
	81-90		2	2	8	2	14
	81-90		2	2	8	2	14
			16,7%	6,9%	17,8%	12,5%	13,2%
			16,7%	6,9%	17,8%	12,5%	13,2%
Total		4	12	29	45	16	106
Total		4	12	29	45	16	106

Anàlisi dels resultats i Comparació amb l'Avaluació Continua

Anàlisi per escoles

Escola 1

Variable	N	Mitj.	Mediana	Mitj.Rc	DesvEst	EEMitj.
RESULTAT	88	44,09	46,50	44,66	13,60	1,45

Variable	Mín	Màx	Q1	Q3
RESULTAT	8,00	69,00	36,00	55,00

Escola 2

Variable	N	Mitj.	Mediana	Mitj.Rc	DesvEst	EEMitj.
RESULTAT	45	38,11	39,00	38,39	15,54	2,32

Variable	Mín	Màx	Q1	Q3
RESULTAT	3,00	67,00	29,00	47,00

Escola 3

Variable	N	Mitj.	Mediana	Mitj.Rc	DesvEst	EEMitj.
RESULTAT	68	39,51	40,50	39,73	13,86	1,68

Variable	Mín	Màx	Q1	Q3
RESULTAT	10,00	64,00	27,25	50,00

Una vegada corregides les proves es va passar un qüestionari al professorat (ANNEX 13, qüestionari) , per tal de comparar els resultats amb els de l'avaluació continuada.

Cada escola va tenir una opinió lleugerament diferent de on establirien la nota per aprovar (55%, 60% i 65%) que fent la mitjana seria d'un 60% del total (42 punts). Aplicant l'error de mesura de 4 punts arrodonits vaig trobar les discrepàncies següents entre la prova i l'avaluació continuada

Escola	Alumnes que passen la prova i no passen per avaluació continua	Alumnes que no passen la prova i passen per avaluació continua
1	**10%**	**6%**
2		**6.7%**
3		**14.7%**

A l'escola 3 s'ha de tenir en compte el fet que més de la meitat dels alumnes havien tingut una professora interina sense experiència en el tipus d'ensenyament d'EOI.

Reflexió metaavaluativa

Les diferències que es troben en les mitjanes dels resultats son estadísticament significatives. Per la qual cosa potser fora bo homogeneïtzar els mètodes de les tasques i els criteris avaluatius mitjançant unes proves fiables i vàlides estandarditzades que els mateixos centres apliquin, com a recurs autoavaluatiu en l'àmbit de l'ensenyament.

Estudis Realitzats

Durant l'any de llicència retribuïda vaig dur a terme el següents estudis:

- **Màster en Direcció i Gestió de Centres Educatius**, organitzat pel Departament de Didàctica i Organització Educativa de la Universitat de Barcelona.
- **Curs d'Anàlisi Estadístic Minitab i Excel**, Departament de Noves Tecnologies, Departament d'Ensenyament.
- Classes Particulars **SPSS**.

Bibliografía Bàsica

1. ALTE. (1998) *Multilingual Glossary of Language Testing Terms.*CUP. Cambridge.

2. Alderson, J. C.; Clapham, C.; Wall, D. (1996) *Language Test Construction and Evaluation CUP* Cambridge

3. Bachman, Lyle. (1990) *Fundamental Considerations in Language Testing.* OUP. Oxford

4. Bachman, Lyle & Palmer, Adrian (1996) *Language Testing in Practice.* OUP. Oxford

5. Council of Europe,(2001) *Marco de Referencia Europeo para el aprendizaje, la Enseñanza y la Evaluación de las Lenguas*

6. Cronbach, L. J. (1951). *"Coeffiecient alpha and the internal structure of tests."* Psychometrika

7. Cronbach, L. J. (1984). Essential of Psychological Testing. Fourth edition. : Harper and Row.Nova York

8. Douglas, D. (2000) *Assessing Languages for Specific Purposes,* CUP. Cambridge.

9. Henning, G. (1987) *A Guide to Language Testing, CUP* Mass.: Newbury House

10. Hernández Sampieri et al, (1999) *Metodología de la Investigación,* Méjico DF

11. Mc Millan, James (2000) *Fundamental Assessment Principles for Teachers and School Administrators*. Virginia Commonwealth University

12. Padua, J (1992) *Técnicas de investigación aplicadas a las ciencias sociales* Fondo de Cultura Económica. México.

13. Pardo Merino, Antonio et al, (2002) *SPSS 11,* Mc Graw Hill. Madrid

14. Pérez, César (2001) Técnicas Estadísticas con SPSS. Prentice Hall. Madrid

15. Peters,C. i Van Voorhis (1940) *Statistical Procedures and their Mathematical Basis* Mc Graw-Hill Co. New York

16. Rul, J. (1995) *La memòria avaluativa del centre educatiu. Un model integral d' avaluació organitzativa i curricular*. Generalitat de Catalunya. Barcelona

17. Rul,J (2001) *Autoavaluació de Centres Educatius. Les proves de Mesura del Rendiment Acadèmic en Grups Petits*. Departament d'Ensenyament. Generalitat de Catalunya. Barcelona

18. Spearman, C. (1910). "*Correlation calculated from faulty data."* British Journal of Psychology

19. Stanley, J. C. (1957). *"K-R 20 as the stepped-up mean item intercorrelation"* 14th Yearbook of the National Council on Measurement in Education. Nova York

20. Stanley, J. C. (1961). "*Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence."* Psychometrika.

RELACIÓ D'ANNEXOS

ANNEX 1, Prova Prèvia

ANNEX 2, notes pretest

ANNEX 3, base dades mtw i spss

ANNEX 4, taula freqüències

ANNEX 5, Score Definitiu

ANNEX 6, Correlacions OD

ANNEX 7, llista de funcions comunes

ANNEX 8, Prova definitiva

ANNEX 9, Notes

ANNEX 10, base dades v1 i v2 definitiva

ANNEX 11, taula freqüències

ANNEX 12, Correlacions

ANNEX 13, qüestionari