

COMMISSION ÉLECTROTECHNIQUE INTERNATIONALE
RAPPORT DE LA CEI

INTERNATIONAL ELECTROTECHNICAL COMMISSION
IEC REPORT

Publication 268-13

Première édition — First edition

1985

Équipements pour systèmes électroacoustiques

Treizième partie: Essais d'écoute des haut-parleurs

Sound system equipment

Part 13: Listening tests on loudspeakers



© CEI 1985

Droits de reproduction réservés — Copyright - all rights reserved

Bureau Central de la Commission Electrotechnique Internationale

3, rue de Varembé

Genève, Suisse

Révision de la présente publication

Le contenu technique des publications de la CEI est constamment revu par la CEI afin d'assurer qu'il reflète bien l'état actuel de la technique.

Les renseignements relatifs à ce travail de révision, à l'établissement des éditions révisées et aux mises à jour peuvent être obtenues auprès des Comités nationaux de la CEI et en consultant les documents ci-dessous:

- **Bulletin de la CEI**
- **Annuaire de la CEI**
- **Catalogue des publications de la CEI**
Publié annuellement

Terminologie

En ce qui concerne la terminologie générale, le lecteur se reportera à la Publication 50 de la CEI: Vocabulaire Electrotechnique International (VEI), qui est établie sous forme de chapitres séparés traitant chacun d'un sujet défini, l'Index général étant publié séparément. Des détails complets sur le VEI peuvent être obtenus sur demande.

Les termes et définitions figurant dans la présente publication ont été soit repris du VEI, soit spécifiquement approuvés aux fins de cette publication.

Symboles graphiques et littéraux

Pour les symboles graphiques, symboles littéraux et signes d'usage général approuvés par la CEI, le lecteur consultera:

- la Publication 27 de la CEI: Symboles littéraux à utiliser en électrotechnique;
- la Publication 617 de la CEI: Symboles graphiques pour schémas.

Les symboles et signes contenus dans la présente publication ont été soit repris des Publications 27 ou 617 de la CEI, soit spécifiquement approuvés aux fins de cette publication.

Publications de la CEI établies par le même Comité d'Etudes

L'attention du lecteur est attirée sur les pages 3 et 4 de la couverture, qui énumèrent les publications de la CEI préparées par le Comité d'Etudes qui a établi la présente publication.

Revision of this publication

The technical content of IEC publications is kept under constant review by the IEC, thus ensuring that the content reflects current technology.

Information on the work of revision, the issue of revised editions and amendment sheets may be obtained from IEC National Committees and from the following IEC sources:

- **IEC Bulletin**
- **IEC Yearbook**
- **Catalogue of IEC Publications**
Published yearly

Terminology

For general terminology, readers are referred to IEC Publication 50: International Electrotechnical Vocabulary (IEV), which is issued in the form of separate chapters each dealing with a specific field, the General Index being published as a separate booklet. Full details of the IEV will be supplied on request.

The terms and definitions contained in the present publication have either been taken from the IEV or have been specifically approved for the purpose of this publication.

Graphical and letter symbols

For graphical symbols, and letter symbols and signs approved by the IEC for general use, readers are referred to:

- IEC Publication 27: Letter symbols to be used in electrical technology;
- IEC Publication 617: Graphical symbols for diagrams.

The symbols and signs contained in the present publication have either been taken from IEC Publication 27 or 617, or have been specifically approved for the purpose of this publication.

IEC publications prepared by the same Technical Committee

The attention of readers is drawn to pages 3 and 4 of the cover, which list IEC publications issued by the Technical Committee which has prepared the present publication.

COMMISSION ÉLECTROTECHNIQUE INTERNATIONALE

RAPPORT DE LA CEI

INTERNATIONAL ELECTROTECHNICAL COMMISSION

IEC REPORT

Publication 268-13

Première édition — First edition

1985

Équipements pour systèmes électroacoustiques

Treizième partie: Essais d'écoute des haut-parleurs

Sound system equipment

Part 13: Listening tests on loudspeakers



© CEI 1985

Droits de reproduction réservés — Copyright - all rights reserved

Aucune partie de cette publication ne peut être reproduite ni utilisée sous quelque forme que ce soit et par aucun procédé, électronique ou mécanique, y compris la photocopie et les microfilms, sans l'accord écrit de l'éditeur.

No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from the publisher.

Bureau Central de la Commission Electrotechnique Internationale

3, rue de Varembe

Genève, Suisse

SOMMAIRE

	Pages
PRÉAMBULE	4
PRÉFACE	4
Articles	
1. Généralités	6
1.1 Domaine d'application	6
1.2 Objet	6
1.3 Définition des termes	6
2. Conditions physiques	8
2.1 Local d'écoute	8
2.2 Emplacement et orientation des haut-parleurs	12
2.3 Position de l'auditeur	14
2.4 Programme	14
2.5 Réglage de niveau	18
2.6 Conditions électriques	20
3. Procédure expérimentale et évaluation	22
3.1 Auditeurs	22
3.2 Durée de l'essai	22
3.3 Echelles d'évaluation	24
3.4 Choix des méthodes d'essai	26
3.5 Instructions et essais préliminaires	28
3.6 Nombre de jugements	30
3.7 Traitement statistique des résultats	30
3.8 Contenu d'un rapport d'essai	30
ANNEXE A — Instructions pour les essais d'écoute	34
ANNEXE B — Comparaison de la longueur d'essai	38
ANNEXE C — Traitement statistique des résultats à partir des essais d'écoute	44
FIGURE 1	70

CONTENTS

	Page
FOREWORD	5
PREFACE	5
Clause	
1. General	7
1.1 Scope	7
1.2 Object	7
1.3 Explanation of terms	7
2. Physical conditions	9
2.1 Listening room	9
2.2 Loudspeaker position and orientation	13
2.3 Listener position	15
2.4 Programme material	15
2.5 Level setting	19
2.6 Electrical requirements	21
3. Experimental procedure and evaluation	23
3.1 Listeners	23
3.2 Test duration	23
3.3 Rating scales	25
3.4 Choice of test procedure	27
3.5 Instruction and preliminary trials	29
3.6 Number of judgements	31
3.7 Statistical treatment of the data	31
3.8 Contents of test report	31
APPENDIX A — Instructions for listening tests	35
APPENDIX B — Comparison of test length	39
APPENDIX C — Statistical treatment of data from listening tests	45
FIGURE 1	71

COMMISSION ÉLECTROTECHNIQUE INTERNATIONALE

ÉQUIPEMENTS POUR SYSTÈMES ÉLECTROACOUSTIQUES

Treizième partie: Essais d'écoute des haut-parleurs

PRÉAMBULE

- 1) Les décisions ou accords officiels de la CEI en ce qui concerne les questions techniques, préparés par des Comités d'Etudes où sont représentés tous les Comités nationaux s'intéressant à ces questions, expriment dans la plus grande mesure possible un accord international sur les sujets examinés.
- 2) Ces décisions constituent des recommandations internationales et sont agréées comme telles par les Comités nationaux.
- 3) Dans le but d'encourager l'unification internationale, la CEI exprime le vœu que tous les Comités nationaux adoptent dans leurs règles nationales le texte de la recommandation de la CEI, dans la mesure où les conditions nationales le permettent. Toute divergence entre la recommandation de la CEI et la règle nationale correspondante doit, dans la mesure du possible, être indiquée en termes clairs dans cette dernière.

PRÉFACE

Le présent rapport a été établi par le Comité d'Etudes n° 84 de la CEI: Equipements et systèmes dans le domaine des techniques audio, vidéo et audiovisuelles. (Anciennement Sous-Comité 29B.)

Le texte de ce rapport est issu des documents suivants:

Règle des Six Mois	Rapport de vote
29B(BC)104	29B(BC)113

Pour de plus amples renseignements, consulter le rapport de vote mentionné dans le tableau ci-dessus.

Les publications suivantes de la CEI sont citées dans le présent rapport:

Publications n°s 268-1 (1968): Equipements pour systèmes électroacoustiques, Première partie: Généralités.
543 (1976): Guide pour l'évaluation subjective par écoute.
651 (1979): Sonomètres.

Autres publications citées:

Norme ISO 389 (1975): Acoustique – Zéro normal de référence pour l'étalonnage des audiomètres à sons purs.
Norme ISO 3382 (1975): Acoustique – Mesurage de la durée de réverbération des auditoriums.

INTERNATIONAL ELECTROTECHNICAL COMMISSION

SOUND SYSTEM EQUIPMENT**Part 13: Listening tests on loudspeakers**

FOREWORD

- 1) The formal decisions or agreements of the IEC on technical matters, prepared by Technical Committees on which all the National Committees having a special interest therein are represented, express, as nearly as possible, an international consensus of opinion on the subjects dealt with.
- 2) They have the form of recommendations for international use and they are accepted by the National Committees in that sense.
- 3) In order to promote international unification, the IEC expresses the wish that all National Committees should adopt the text of the IEC recommendation for their national rules in so far as normal conditions will permit. Any divergence between the IEC recommendation and the corresponding national rules should, as far as possible, be clearly indicated in the latter.

PREFACE

This report has been prepared by IEC Technical Committee No. 84: Equipment and Systems in the Field of Audio, Video and Audiovisual Engineering. (Formerly Sub-Committee 29B.)

The text of this report is based upon the following documents:

Six Months' Rule	Report on Voting
29B(CO)104	29B(CO)113

Further information can be found in the Report on Voting indicated in the table above.

The following IEC publications are quoted in this report:

Publications Nos. 268-1 (1968): Sound System Equipment, Part 1: General.
543 (1976): Informative Guide for Subjective Listening Tests.
651 (1979): Sound Level Meters.

Other publications quoted:

ISO Standard 389 (1975): Acoustics – Standard Reference Zero for the Calibration of Pure-tone Audiometers.
ISO Standard 3382 (1975): Acoustics – Measurement of Reverberation Time in Auditoria.

ÉQUIPEMENTS POUR SYSTÈMES ÉLECTROACOUSTIQUES

Treizième partie: Essais d'écoute des haut-parleurs

1. Généralités

1.1 *Domaine d'application*

Les recommandations figurant dans le présent rapport s'appliquent aux haut-parleurs destinés à des usages et ambiances domestiques. Bien que spécifiquement conçues pour les haut-parleurs en tant qu'éléments distincts de systèmes électroacoustiques, les procédures sont également applicables, avec des modifications mineures, à d'autres dispositifs tels que les récepteurs de radio et les appareils de télévision.

1.2 *Objet*

Le présent rapport donne des recommandations pour la mise en œuvre, la validité et l'interprétation des essais d'écoute des haut-parleurs.

Les essais décrits dans ce rapport doivent être effectués dans un local possédant des caractéristiques acoustiques semblables à celles d'une salle de séjour «moyenne». Les recommandations concernent les dimensions du local, le traitement et les mesures acoustiques, la disposition des haut-parleurs et des auditeurs ainsi que les conditions ambiantes.

On y décrit des procédures expérimentales comportant des recommandations sur le choix du programme musical et parlé ainsi que sur le traitement et la présentation des résultats.

1.3 *Définition des termes*

Dans le présent rapport, certains termes particuliers sont utilisés avec le sens suivant:

Stimulus

Reproduction d'un programme donné sur un haut-parleur donné.

Programme

Une brève séquence musicale ou parlée (d'environ 30 s) présentée sans interruption par l'un des haut-parleurs à un moment donné de l'essai d'écoute.

Répétition

Écoute renouvelée du même stimulus au cours de l'essai, destinée à augmenter la fiabilité des évaluations.

Fiabilité

- a) La fiabilité individuelle («pour un auditeur donné») se rapporte à une concordance entre des évaluations répétées d'un même auditeur sur un même stimulus.
- b) La fiabilité interindividuelle («entre plusieurs auditeurs») se rapporte à une concordance entre les évaluations d'un même stimulus par les différents auditeurs.

Interaction

L'interaction entre deux variables signifie que les effets de l'une d'elles sont influencés par le niveau de l'autre. Appliquée aux évaluations dans un essai d'écoute, une interaction entre les haut-parleurs et les programmes peut signifier que les écarts de notation entre deux (ou plusieurs) haut-parleurs sont différents pour des programmes différents.

SOUND SYSTEM EQUIPMENT

Part 13: Listening tests on loudspeakers

1. General

1.1 Scope

The recommendations in this report apply to loudspeakers intended for household systems and environments. Although specifically designed for loudspeakers that are separate sound system components the procedures, with minor changes, are applicable also to other devices such as radio receivers and television sets.

1.2 Object

This report gives recommendations for the setting-up, performance and evaluation of listening tests on loudspeakers.

The tests described in this report are to be performed in a room having acoustical characteristics similar to those of an “average” living room. Recommendations about the room size, acoustical treatment and measurements, arrangement of loudspeakers and listeners, and environmental conditions are given.

Experimental procedures are described, including recommendations on the choice of music and speech programme material and the processing and presentation of the final data.

1.3 Explanation of terms

In this report, a number of special terms are used with the following meanings:

Stimulus

Reproduction of a certain programme section over a certain loudspeaker.

Programme section

A short piece of music or speech (approximately 30 s duration) that is presented without interruption over one loudspeaker at a time in a listening test.

Replication

Repeated application of the same stimulus in the test in order to increase the reliability of the ratings.

Reliability

- a) Intra-individual (“within listener”) reliability refers to the agreement between a certain listener’s repeated ratings of the same stimulus.
- b) Inter-individual (“between listeners”) reliability refers to the agreement between different listeners’ ratings of the same stimulus.

Interaction

An interaction between two variables means that the effects of one variable are different at different levels of the other variable. Applied to the ratings in a listening test, an interaction between loudspeakers and programmes may mean that the differences between the ratings of two (or more) loudspeakers are different for different programmes.

2. Conditions physiques

2.1 Local d'écoute

L'idéal serait que le local d'écoute simule un environnement d'écoute domestique typique de la zone géographique où les résultats d'essai sont applicables. Le présent rapport décrit les principes connus pour la conception d'un bon local d'écoute, afin d'aider l'utilisateur à réaliser le meilleur environnement d'écoute possible tout en tenant compte des contraintes dues à sa situation et au contexte local. Certaines recommandations plus spécifiques sont données pour définir «de façon internationale» un local type d'écoute destiné à être utilisé lorsque les résultats des essais doivent avoir une application aussi vaste que possible.

2.1.1 Dimensions du local

Dans les locaux de dimensions domestiques, la qualité du son aux basses fréquences est sensiblement influencée par la distribution des fréquences propres du local et par l'emplacement des haut-parleurs et des auditeurs dans le réseau des ondes stationnaires.

Etant donné que la distribution des fréquences propres à un local rectangulaire ayant des cloisons planes parfaitement réfléchissantes est définie par le rapport des dimensions du local (hauteur: longueur: largeur), le choix des proportions optimales est une préoccupation naturelle. En pratique, toutefois, de telles considérations doivent être modifiées par le fait que la plupart des parois des locaux ne sont ni parfaitement planes ni parfaitement réfléchissantes aux basses fréquences.

Le fait que la position du haut-parleur et de l'auditeur affecte le couplage acoustique dans des modes propres du local signifie que des colorations perceptibles à l'oreille peuvent être provoquées par le renforcement fortuit d'un certain nombre de modes résonnants, dans un local qui serait acceptable dans d'autres conditions. Il semble même que le mode axial revête une importance particulière à cet égard. Il en résulte que la réponse en fréquence du canal de transmission que constitue le local d'écoute dépendra de plusieurs facteurs dont tous ne seront pas connus d'avance.

Il est donc recommandé d'inclure dans la qualification définitive du local d'écoute un examen de la transmission acoustique au moyen de mesures de réponse en fréquences glissantes. (Des mesures avec une résolution moins fine peuvent être effectuées en utilisant un bruit rose filtré en bandes de tiers d'octave.) Il convient de placer un bon haut-parleur à large bande successivement à tous les emplacements d'essai du haut-parleur, un microphone de mesure étant placé successivement à tous les emplacements des têtes des auditeurs. On devrait explorer toutes les combinaisons possibles et effectuer les ajustements de position pour éviter les quelques situations qui posent réellement des problèmes. Des aberrations importantes de la réponse en fréquence sont susceptibles de se produire à des fréquences inférieures à environ 250 Hz.

Les dimensions du local d'écoute imposent aussi bien les fréquences propres que leur densité dans le domaine de fréquences. Les locaux spacieux offrent une possibilité plus grande d'obtenir une réponse en fréquence uniforme, mais ces dimensions comportent, en pratique, une limite supérieure.

Les valeurs données à la figure 1, page 70, indiquent les limites acceptables pour les dimensions du local. A l'intérieur de ces limites, les dimensions recommandées ci-dessous définissent ce qui est considéré comme un bon local d'écoute dans la gamme des dimensions typiques pour de nombreuses régions géographiques.

	Dimension recommandée
Volume	80 m ³
Hauteur	2,80 m
Longueur	6,70 m
Largeur	4,20 m

Note. — Les dimensions recommandées pour le local d'écoute rentrent dans le cadre défini par la Publication 543 de la CEI: Guide pour l'évaluation subjective par écoute.

2. Physical conditions

2.1 Listening room

Ideally, the listening room should simulate a typical household listening environment in the geographical region where the test results apply. This report describes known principles of good listening room design to assist the user in approaching an optimum listening environment as closely as possible within the constraints of his situation and local context. Some more specific recommendations are given to describe an “international standard” listening room for use when test results are required to have the widest possible application.

2.1.1 Room dimensions

In rooms of household size, sound quality at low frequencies is significantly influenced by the frequency distribution of room eigentones and by the placement of loudspeakers and listeners in the pattern of standing-waves.

Because the distribution of eigentones in a rectangular room with flat, perfectly-reflecting walls is described by the ratio of room dimensions (height: length: breadth), the choice of optimum proportions is a natural consideration. In practice, however, such considerations must be modified by the fact that most room boundaries are neither perfectly-flat nor perfectly-reflecting at low frequencies.

The fact that the positioning of loudspeaker and listener affects the acoustical coupling into individual room modes means that audible colorations can be caused by the fortuitous enhancement of a small number of resonant modes in an otherwise acceptable room. There is even some evidence to suggest that axial modes may be of particular importance in this respect. As a result the frequency response of the transmission path through the listening room will be dependent upon several factors, not all of which are predictable.

It is therefore recommended that the final qualification of the listening room include an inspection of the acoustical transmission paths by means of swept-tone frequency-response measurements. (Measurements with poorer resolution may be made using pink noise filtered in 1/3 octave bands.) A good, broad-band loudspeaker should be positioned, in turn, at all loudspeaker test locations with a measuring microphone placed, in turn, at all listener head locations. All combinations should be explored and positional adjustments made to avoid those situations that clearly present difficulties. Large aberrations in frequency response tend to occur at frequencies below about 250 Hz.

The size of the room dictates both the specific eigentone frequencies and their density in the frequency domain. Large rooms offer greater potential for uniform frequency response but there is a practical upper limit on size.

The values given in Figure 1, page 71, indicate the acceptable ranges of room dimensions. Within those ranges, the recommended dimensions given below describe what is believed to be a good listening room falling within the range of typical dimensions for many geographical regions.

	Recommended dimension
Volume	80 m ³
Height	2.80 m
Length	6.70 m
Breadth	4.20 m

Note. — The recommended listening room dimensions come within the framework of those defined in IEC Publication 543: Informative Guide for Subjective Listening Tests.

2.1.2 Temps de réverbération

Les sons réfléchis et réverbérants dans le local d'écoute ont une influence majeure sur la qualité sonore globale et sur la perception stéréophonique. La quantité et le type des matériaux absorbants acoustiques et l'emplacement de ces matériaux imposent le temps de réverbération, l'uniformité du champ acoustique réverbérant (degré de diffusion) et l'interaction entre le haut-parleur et les proches parois du local. Dans les ambiances domestiques, des facteurs régionaux et économiques dans l'architecture et le décor jouent un rôle important dans la définition de ce qui est «typique» à cet égard.

Techniquement, le temps de réverbération T ainsi que la régularité de sa variation en fonction de la fréquence sont importants. Le temps de réverbération doit être mesuré conformément à la Norme ISO 3382, à intervalles d'un tiers d'octave. Les mesures doivent être effectuées en l'absence d'auditeurs.

Entre 250 Hz et 4 000 Hz, la valeur moyenne de T doit être comprise dans l'intervalle de 0,3 s à 0,6 s. Dans cette gamme de fréquences, les mesures individuelles de T ne doivent pas s'écarter de la valeur moyenne de plus de 25%. Au-dessous de 250 Hz et au-dessus de 4 000 Hz, T peut s'écarter de plus de 25% de sa valeur moyenne pour la fréquence médiane, mais ne devra pas dépasser 0,8 s au-dessous de 250 Hz.

Là où les résultats des essais d'écoute sont d'un intérêt débordant le plan régional ou particulier, on recommande que la valeur moyenne de T à la fréquence médiane soit de $0,4 \pm 0,05$ s.

En ce qui concerne la disposition matérielle, le local devra avoir un plafond qui soit principalement réfléchissant et un plancher qui soit presque entièrement recouvert de moquette. Les cloisons situées derrière et à proximité immédiate sur les côtés des haut-parleurs et le plafond, sur un tiers environ de la longueur du local, devront être pour la plupart réfléchissants pour le son. Les échos flottants devront être supprimés au moyen de matériaux absorbants et diffusants, disposés de telle manière qu'une surface réfléchissante ne se trouve pas en face d'une autre surface réfléchissante. La paroi derrière les auditeurs devra être traitée pour éviter de fortes réflexions cohérentes, en particulier, aux moyennes et hautes fréquences.

L'intérêt du local sera considérablement augmenté en y introduisant des panneaux absorbants mobiles. Ceux-ci peuvent être utilisés pour amortir acoustiquement les parois derrière ou à côté des haut-parleurs, ou ils peuvent être répartis autour de la zone d'écoute selon le souhait du constructeur ou de l'expérimentateur. Une méthode commode pour obtenir cette souplesse consiste dans l'utilisation de tentures tombant du plafond au plancher, en deux ou plusieurs parties, et dont la longueur totale est égale à peu près à la largeur du local. Les tentures peuvent être suspendues à un rail, à proximité des parois, et s'étendre au moins à mi-chemin autour du local. Le tissu des tentures devra être choisi pour sa bonne absorption acoustique et devra être placé à l'endroit destiné à un essai d'écoute spécifique pendant les mesures du temps de réverbération.

Note. — Il convient d'attirer particulièrement l'attention sur le fait que la quantité et la disposition des matériaux absorbants et diffusants dans le local sont susceptibles d'influencer les résultats des essais stéréophoniques. La perception de localisation de l'image sonore par les auditeurs et leurs impressions d'espace peuvent être affectées d'une façon différente par l'intensité et la direction des sons réfléchis et réverbérants dans le local d'écoute.

2.1.3 Conditions ambiantes

La fiabilité des résultats peut être affectée par des conditions ambiantes extrêmes; en conséquence, les conditions d'écoute devront être aussi confortables que possible. Pour les essais, on doit observer les prescriptions suivantes:

Température ambiante:	18°C à 35°C, de préférence 20°C
Humidité relative:	25% à 75%
Pression atmosphérique:	86 kPa à 106 kPa (860 mbar à 1 060 mbar).

Le niveau du bruit de fond mesuré dans la zone d'écoute du local, en l'absence d'auditeurs, doit être inférieur à 35 dB (pondération A, lente).

2.1.2 Reverberation time

Reflected and reverberant sounds in the listening room have a major influence on overall sound quality and stereo perception. The amount and type of acoustical absorbing materials and the placement of the materials dictate the reverberation time, the uniformity of the reverberant sound field (degree of diffusion) and the interaction of the loudspeaker with adjacent room boundaries. In the household environment, regional and economic factors in structure and décor are significant factors in defining what is "typical" in these respects.

Technically, both the reverberation time, T , and the regularity of its variation as a function of frequency are important. Reverberation time shall be measured according to ISO Standard 3382 at 1/3 octave intervals, without listeners present.

Between 250 Hz and 4000 Hz the average value of T shall be within the range 0.3 s to 0.6 s. In that frequency range individual measurements of T should not deviate by more than 25% from the average value. Below 250 Hz and above 4000 Hz, T is permitted to deviate by more than 25% from the middle-frequency average value, but should not exceed 0.8 s below 250 Hz.

Where the results of the listening tests are of more than regional or other specific interest, it is recommended that the middle-frequency average, T , be 0.4 ± 0.05 s.

In physical arrangement the room should have a ceiling that is mostly sound reflecting and a floor that is mostly carpeted. The walls behind and immediately to the sides of the loudspeakers and the ceiling for approximately one-third of the room length should be mostly sound reflecting. Flutter echoes should be suppressed by means of sound absorbing and scattering objects arranged in such a manner that no reflecting surface is located opposite another parallel reflecting surface. The wall behind the listeners should be treated to prevent strong coherent reflections, particularly at middle and high frequencies.

The usefulness of the room will be substantially increased by the inclusion of some moveable sound-absorbing surfaces. These can be used to damp the walls acoustically, behind or beside the loudspeakers, or they can be distributed around the listening area, as the manufacturer or the experimenter requires. A convenient method of achieving this flexibility is to use floor-to-ceiling length drapes, in two or more sections, having a total length approximately equal to the width of the room. The drapes can be suspended from a rail near to the perimeter of the room and extending at least halfway around the room. The drapery material should be chosen for high sound absorption and should be in the place intended for a specific listening test during the reverberation-time measurements.

Note. — It should be noted especially that the quantity and arrangement of sound absorbing and scattering materials in the room are likely to influence the results of stereophonic tests. The listeners' perceptions of sound-image localization and their impressions of space may be differently affected by the strength and direction of reflected and reverberant sounds in the listening room.

2.1.3 Climatic conditions

The reliability of results may be affected by extremes of climatic conditions, therefore listening conditions should be as comfortable as possible. For test purposes the following requirements shall be met:

Ambient temperature:	18°C to 35°C, preferably 20°C
Relative humidity:	25% to 75%
Air pressure:	86 kPa to 106 kPa (860 mbar to 1060 mbar).

The background noise level measured in the listening area of the room, in the absence of listeners, shall be below 35 dB (A-weighted, slow).

2.2 *Emplacement et orientation des haut-parleurs*

L'emplacement d'un haut-parleur par rapport aux limites adjacentes du local (plancher, cloisons, plafond) affecte aussi bien l'aspect spectral que l'aspect temporel des sons se propageant dans le local d'écoute. En conséquence, le choix de l'emplacement du haut-parleur est un facteur important pour les résultats des essais d'écoute.

En général, il convient de placer les haut-parleurs conformément aux instructions des constructeurs. Toutefois, la plupart des constructeurs ne donnent (si toutefois ils en donnent) que des conseils généraux concernant l'emplacement optimal. Par conséquent, il est recommandé que les haut-parleurs soient écoutés dans plusieurs emplacements différents. C'est l'une des raisons pour lesquelles il est utile d'effectuer, en général, deux sortes d'essais: les essais monophoniques bien adaptés à l'appréciation des qualités sonores fondamentales et dans lesquels les emplacements du haut-parleur peuvent être facilement changés, et les essais stéréophoniques qui comprennent des appréciations des qualités complémentaires du son reproduit, et dans lesquels les emplacements des haut-parleurs sont imposés de façon rigoureuse.

Sauf instructions spécifiques contraires de la part du constructeur, l'axe de référence de chaque haut-parleur doit être horizontal, situé à 1,25 m au-dessus du plancher et orienté vers un point proche du centre de l'emplacement des sièges des auditeurs se trouvant sur la ligne médiane du local. Les haut-parleurs doivent être orientés de façon que le cône d'émission ait son plus grand angle dans le plan horizontal passant par l'axe de référence.

Les haut-parleurs en essai doivent être rendus invisibles aux auditeurs au moyen d'un écran acoustiquement transparent séparant le haut-parleur des zones d'écoute (voir figure 1, page 70). L'effet de l'écran sur la réponse en fréquence des haut-parleurs doit être inférieur à 1 dB à n'importe quelle audiofréquence (mesuré avec l'écran placé à mi-chemin entre un haut-parleur et un microphone distant de 2 m).

2.2.1 *Essais monophoniques*

Les haut-parleurs qui, par leur conception (par exemple pavillons d'angle) ou pour d'autres considérations, doivent être placés conformément à une spécification du constructeur, devront être, bien entendu, écoutés de cette façon. Toutefois, la plupart des haut-parleurs peuvent être placés dans une partie du local d'écoute, indiquée dans la figure 1 comme «zone de haut-parleurs», qui est délimitée de façon que le haut-parleur soit disposé à au moins 1 m des parois latérales et à au moins 0,70 m de la paroi arrière. Ces distances sont mesurées à partir du point de référence (intersection de l'axe de référence et du plan frontal du haut-parleur).

Les différents haut-parleurs soumis à une comparaison monophonique devront être séparés par 0,50 m au moins. Les emplacements choisis devront être marqués et les haut-parleurs devront être permutés, si possible, au cours des essais afin de ne pas influencer les résultats globaux en corrélatant certains haut-parleurs avec des emplacements particuliers.

Le rapport d'essai devra indiquer les emplacements des haut-parleurs utilisés pour chaque séance d'écoute.

Note. — Les dispositifs mécaniques qui remplacent les différents haut-parleurs en un même emplacement du local entraînent un manque de souplesse dans la disposition des haut-parleurs, mais peuvent être utiles dans certaines applications.

2.2.2 *Essais stéréophoniques*

L'objectif principal des essais d'écoute stéréophonique est d'évaluer l'aptitude des haut-parleurs à simuler des images sonores localisées en azimut, site et distance, ainsi que l'ambiance acoustique d'une prestation originale. Etant donné que plus d'une paire de haut-parleurs ne peut occuper à la fois les mêmes emplacements, les évaluations de cette nature sont susceptibles d'être perturbées par la commutation des positions de la source. Un dispositif mécanique de substitution peut être une solution satisfaisante pour être utilisé avec plusieurs configurations courantes de haut-parleurs.

2.2 Loudspeaker position and orientation

The location of a loudspeaker with respect to the immediately adjacent room boundaries (floor, walls and ceiling) affects both the spectral and temporal aspects of sounds propagated into the listening room. The choice of loudspeaker location, therefore, is a factor of importance in the results of the listening tests.

In general, loudspeakers should be placed according to the recommendations of the manufacturers. However, most manufacturers offer only general guidance, if any, about optimum placement. Consequently, it is recommended that loudspeakers be auditioned in several different room locations. This is one reason why it is usually advisable to conduct two kinds of tests: monophonic tests, well suited to judgements of basic sound qualities and in which loudspeaker locations can be easily changed, and stereophonic tests that involve judgements of additional attributes of the reproduced sound and in which loudspeaker locations are rigorously controlled.

Unless the manufacturer gives specific indications to the contrary, the reference axis of each loudspeaker shall be horizontal, 1.25 m above the floor, and pointed towards a point near the centre of the listener seating area on the midline of the room. Loudspeakers shall be oriented so that the broadest dispersion of sound from the loudspeaker occurs in the horizontal plane passing through the reference axis.

The loudspeakers under test shall be rendered invisible to the listeners by an acoustically transparent screen separating the loudspeaker and listening areas (see Figure 1, page 71). The effect of the screen on the frequency response of the loudspeakers shall be less than 1 dB at any audio-frequency (measured with the screen midway between a loudspeaker and microphone spaced at 2 m).

2.2.1 Monophonic tests

Loudspeakers that by design (e.g. corner horns) or other considerations must be located according to a manufacturer's specification should, of course, be auditioned accordingly. The majority of loudspeakers, however, can be accommodated within a portion of the listening room, designated in Figure 1 as the "loudspeaker area", that is delineated by the restriction that the loudspeaker shall be at least 1 m from the side walls and at least 0.70 m from the back wall. These distances are measured to the reference point (the intersection of the reference axis and the front plane of the loudspeaker).

The different loudspeakers in a monophonic comparison should be separated by at least 0.50 m. The chosen locations should be noted and, if possible, the loudspeakers should be interchanged during the tests so as to avoid biasing the overall results by associating particular loudspeakers with particular room positions.

The test report should include a description of the loudspeaker locations used for each listening session.

Note — Mechanical devices that substitute the different loudspeakers in the same room location suffer through a lack of flexibility in loudspeaker placement, but may be useful in some applications.

2.2.2 Stereophonic tests

The main purpose of stereophonic listening tests is to judge the ability of loudspeakers to reproduce the illusion of sound images localized in azimuth, elevation and depth, as well as the acoustical ambiance of an original performance. Since more than one pair of loudspeakers cannot, at one time, occupy the same room position, judgements of this type tend to be confounded by the switching of source locations. A mechanical substitution device may be a satisfactory solution for use with many popular loudspeaker configurations.

Par conséquent, dans des conditions normales, il est recommandé d'effectuer les essais d'écoute stéréophonique en n'utilisant qu'une paire de haut-parleurs à la fois. Les essais comparatifs qui ne prévoient pas de substitution des positions des haut-parleurs devront être effectués avec précaution.

Les haut-parleurs devront être séparés d'au moins 2 m. Dans des locaux plus petits, il peut être nécessaire de placer les haut-parleurs à moins de 0,50 m des parois latérales pour obtenir la distance nécessaire à un bon relief stéréophonique. L'auditeur le plus proche devra être éloigné d'au moins 2 m, mesurés perpendiculairement à la ligne imaginaire reliant les haut-parleurs. En général, la distance d'écoute minimale devra être égale à 1 fois (voire 0,8 fois) la distance entre les haut-parleurs, c'est-à-dire que les haut-parleurs devront être vus sous un angle de 55° à 65° depuis l'emplacement de l'auditeur le plus proche. Pour une bonne reproduction stéréophonique, la disposition des haut-parleurs et des auditeurs devra être symétrique par rapport à l'axe principal du local.

Sauf instructions spécifiques contraires de la part du constructeur, les haut-parleurs devront être orientés de façon que les axes de référence se croisent sur l'axe principal du local à un point proche du centre de la ligne des auditeurs.

2.3 Position de l'auditeur

Les auditeurs doivent être assis dans la zone du local spécifiée dans la figure 1, page 70. Afin de compenser en moyenne les propriétés directionnelles des haut-parleurs et les propriétés de couplage acoustique du local, il est avantageux d'utiliser plusieurs positions d'écoute au cours d'un essai. Dans les essais stéréophoniques, il est essentiel que les auditeurs soient alignés dans le plan de symétrie de la ligne de haut-parleurs. Du premier au dernier rang de la file, les auditeurs devront être progressivement placés plus haut de façon à donner à chaque tête une «vue» acoustique claire de tous les haut-parleurs en essai. Dans les essais monophoniques, il suffit qu'une seule position d'auditeur soit sur l'axe du local, les autres pouvant être placées de façon aléatoire à l'intérieur de la zone suggérée.

Chaque auditeur devra être placé à 0,40 m au moins des parois latérales et à 1 m au moins du mur arrière. La distance entre auditeurs devra être de 0,60 m au moins.

Les chaises destinées aux auditeurs devront avoir des dossiers ne dépassant pas la hauteur de l'épaule pour éviter une interférence acoustique dans la perception normale du champ acoustique total du local.

Bien qu'il puisse y avoir un nombre indéterminé de places d'auditeurs, le nombre d'auditeurs à chaque séance devra être limité. Les essais les plus critiques peuvent se faire avec un seul auditeur, occupant les différentes places à des moments différents. Dans les locaux spacieux, cinq auditeurs au maximum peuvent être présents en même temps.

Avec plusieurs auditeurs, il est difficile d'éviter une communication réciproque, qui encourage des réponses collectives plutôt que des réponses individuelles. La communication visuelle peut être réduite par une séparation visuelle des auditeurs en utilisant des écrans transparents au son. Dans tous les cas, les auditeurs devront recevoir des instructions spécifiques soulignant l'importance des réponses individuelles.

2.4 Programme

La sélection de séquences musicales ou parlées pour un essai d'écoute peut avoir une très grande influence sur les résultats. Il existe souvent des interactions entre des haut-parleurs particuliers et des genres particuliers de programme à un point tel que les classements des divers ensembles dépendent des séquences musicales ou parlées présentées. En conséquence, le programme devra être traité comme une variable indépendante distincte.

Under normal circumstances, therefore, it is recommended that stereophonic listening tests be conducted using one pair of loudspeakers at a time. Comparative tests that do not involve positional substitution of loudspeakers should be treated with caution.

The loudspeakers should be separated by at least 2 m. In smaller rooms it may be necessary to move the loudspeakers to within 0.5 m of the side walls to achieve the separation necessary for good stereophonic perspective. The nearest listener should be no less than 2 m, measured perpendicularly, from an imaginary line joining the loudspeakers. In general, the minimum listening distance should be 1.0 to 0.8 times the distance between the loudspeakers, i.e. the loudspeakers should subtend an angle of 55° to 65° at the nearest listener position. For good stereophonic reproduction the arrangement of loudspeakers and listeners should be symmetrical about the major axis of the room.

Unless the manufacturer gives specific instructions to the contrary, the loudspeakers should be oriented so that the reference axes intersect on the major axis of the room at a point near the centre of the line of listeners.

2.3 *Listener position*

Listeners shall be seated in the area of the room specified in Figure 1, page 71. For the purpose of averaging over the directional properties of the loudspeakers and the acoustical coupling properties of the room it is advantageous to use several listening positions in the course of a test. In stereophonic tests it is essential that listeners be arranged in a line on the axis of symmetry of the loudspeaker array. Listeners from front-to-rear of the line arrangement should be progressively elevated to give each head a clear acoustical "view" of all the loudspeakers under test. In monophonic tests only one listener position need be on the room axis; the others can be placed at random within the suggested area.

Each listener should be placed at least 0.40 m from the side walls and at least 1 m from the back wall. The distance between listeners should be at least 0.60 m.

Chairs for listeners should have backs no higher than shoulder level to prevent acoustical interference with normal perception of the total room sound field.

Although there can be any number of listener positions the number of listeners at any session should be limited. The most critical tests may involve only single listeners, occupying the different positions at different times. In large rooms a maximum of five listeners may be used simultaneously.

With multiple listeners it is difficult to avoid mutual communication that encourages group rather than individual responses. Visual communication can be reduced by a visual separation of listeners using acoustically transparent screens. In any event, listeners should be given specific instructions emphasizing the importance of independent responses.

2.4 *Programme material*

The selection of music or speech passages for programme material in a listening test can have a very strong effect on the results of the test. Often there are interactions between specific loudspeakers and particular types of programme to such an extent that the ratings of the various systems will depend on the passages of music or speech presented. The programme material should therefore be treated as a separate independent variable.

Les programmes devront être différents les uns des autres de façon à faire ressortir différentes caractéristiques (dimensions) de la qualité sonore perçue des haut-parleurs.

Note. — S'il était possible d'isoler des caractéristiques spécifiques de la qualité acoustique perçue, chaque programme pourrait être choisi pour mesurer seulement cette caractéristique. Un ensemble de programmes, choisis de façon arbitraire, est susceptible d'être partiellement redondant, c'est-à-dire que plusieurs programmes peuvent révéler les mêmes caractéristiques et, par conséquent, augmenter la longueur de l'essai sans ajouter de nouvelles informations.

Une telle redondance peut être réduite, en examinant les résultats d'essais à partir de divers programmes et en éliminant certains d'entre eux qui offrent des résultats essentiellement identiques. Du fait de cette procédure, l'essai d'écoute devient plus efficace.

Il existe aussi le risque que l'essai ne soit pas complet, ce qui veut dire que certaines caractéristiques de qualité acoustique ne sont l'objet d'aucun programme. L'examen des critères de perception indiqués au paragraphe 3.3 peut donner quelques indications pour le choix des programmes qui, dans l'ensemble, recouvrent la plupart des caractéristiques de la qualité acoustique perçue.

Un programme devra comporter au moins cinq programmes différents. Il est recommandé de consacrer un programme à la parole prononcée de préférence par une voix d'homme au niveau d'une simple conversation. Une partie des instructions d'essai enregistrées peut être utilisée à cet effet. La parole doit être enregistrée dans une chambre anéchoïque sans aucune autre source sonore perceptible que la voix de l'orateur. Le microphone doit être omnidirectionnel, positionné dans le même plan horizontal que la bouche de l'orateur, à une distance de 0,50 m de ce dernier, et déporté d'un côté, la ligne reliant le microphone et la bouche formant un angle de 30° avec le plan sagittal de la tête de l'orateur.

Un autre programme devra comprendre un orchestre symphonique complet, jouant fortissimo, ce qui, comme on le sait, différencie considérablement les haut-parleurs l'un de l'autre. Les autres programmes musicaux devront contenir les combinaisons de petits nombres d'instruments, tels que :

- piano, violon et violoncelle,
- bois et cordes,
- cuivres,
- instruments à percussion
- chant avec accompagnement musical,
- chœurs.

Les programmes musicaux susmentionnés sont destinés à vérifier la précision de la reproduction sonore et, en conséquence, sont tirés des répertoires classiques ou de jazz qui peuvent être comparés à un «original» acoustique, par exemple dans une salle de concert. Les enregistrements de musique légère sont susceptibles d'être des créations de studio, comportant une compression artificielle, des colorations et des effets spatiaux. Certains types de musique légère tels que rock et disco, lorsqu'ils sont reproduits à un fort niveau, imposent aux haut-parleurs des contraintes que la musique classique ne permet pas d'atteindre. En évaluant les matériels pour le grand public, il est important d'inclure quelques essais de leur aptitude à reproduire, d'une façon satisfaisante, ce genre de musique.

Les programmes musicaux devraient présenter une grande variété de genres musicaux et de niveaux sonores, mais les variations de timbre et de niveau sonore à l'intérieur de chaque programme devront être aussi faibles que possible.

Le programme enregistré doit avoir une haute qualité technique. Idéalement, la chaîne d'enregistrement et de reproduction devrait être exempte de distorsions d'amplitude ou de fréquence et les bruits de fond devraient être réduits au minimum. Des enregistrements originaux, faits sous un contrôle rigoureux, sont souhaitables, mais les enregistrements commerciaux, soigneusement sélectionnés, sont acceptables. Par ailleurs, aucune compression d'amplitude automatique ou manuelle ne doit être utilisée à une extrémité de la chaîne sans une expansion correspondante à l'autre extrémité.

The programme sections should differ from each other in a way that brings out different aspects (dimensions) of the perceived sound quality of the loudspeakers.

Note. — If it should prove feasible to isolate specific aspects of perceived sound quality, each section of programme material could be selected to measure that aspect alone. A collection of programmes chosen arbitrarily is liable to be partially redundant, since several programmes may have the same characteristics and hence increase the test length without adding any new information.

Such redundancy can be reduced by examining the test results from a variety of programmes and eliminating some of the sections that yield essentially identical results. By this procedure the listening test is made more efficient.

There is also a risk that the test may be incomplete, which means that certain aspects of sound quality are not covered by any of the programmes. Inspection of the perceptual dimensions given in Sub-clause 3.3 may give certain clues for selecting programmes which together will extend over most aspects of perceived sound quality.

At least five different programme sections should be included. It is recommended that one section consists of speech, preferably a male voice at normal conversation level. Part of the recorded test instruction can be used for this purpose. The speech shall be recorded in an anechoic room with no perceptible source of sound other than the speaker. The microphone shall be of the omnidirectional type, placed in the same horizontal plane as the speaker's mouth at a distance of 0.50 m from the latter and to one side, the line joining microphone and mouth forming an angle of 30° with the sagittal plane of the speaker's head.

Another section should be a full symphony orchestra playing fortissimo, which is known to differentiate strongly between loudspeakers. Other music sections should contain combinations of small numbers of instruments, such as:

- piano, violin and cello,
- woodwind and strings,
- brass instruments,
- percussion instruments,
- solo voice and instrumental accompaniment,
- chorus of voices.

The preceding music sections are intended to test the accuracy of sound reproduction and, therefore, they rely on music of the classical or jazz repertoire which can be compared with an acoustical "original", for example in the concert hall. Recordings of popular music tend to be studio creations, incorporating artificial compression, colorations and spatial effects. Certain types of popular music, such as rock and disco, when reproduced at high sound levels, stress speakers in ways that classical music cannot. In evaluating products for the general public it is important to include some tests of their ability to satisfactorily reproduce this kind of musical material.

The music sections should demonstrate a variety of musical types and sound levels, but the variations of timbre and loudness within each section should be as small as possible.

The recorded programme material shall be of high technical quality. Ideally there should be no frequency or amplitude distortions in the recording and reproduction chain, and background noises should be minimized. Original recordings made under strict control are desirable but carefully-selected commercial recordings are acceptable. Furthermore, no manual or automatic amplitude compression shall be used at one end of the chain without the corresponding expansion at the other end.

Dans les essais monophoniques, le programme devra être sélectionné avec un soin particulier pour s'assurer que des colorations audibles ne se produisent pas du fait de la combinaison de deux canaux stéréophoniques en un seul. Une telle prescription élimine de nombreux enregistrements stéréophoniques commerciaux.

Note. — Etant donné que des variations incontrôlées existent dans les enregistrements commerciaux (le nombre de microphones utilisés, leurs types et positionnement, corrections, réverbération artificielle, etc.), il est parfois opportun d'utiliser des programmes musicaux redondants provenant d'origines variées (différentes sociétés phonographiques, studios, ingénieurs, etc.). La musique et ses sources seront citées dans le rapport d'essai. La conception d'une bande de programme normalisé reconnu sur le plan international est à l'étude. Une telle bande faciliterait la comparaison des essais d'écoute effectués à des endroits différents.

Lorsque le programme musical doit être utilisé pour essayer autre chose que des éléments d'une chaîne haute fidélité (par exemple, récepteurs pour émissions en m.a.), une limitation de bande est autorisée. La fonction de transfert précise doit être spécifiée dans ces cas.

Les programmes devront avoir environ la même durée, entre 20 s et 40 s. Pour des comparaisons par paires, des programmes identiques sont utilisés dans une paire. Les programmes musicaux devront former, autant que possible, des phrases musicales complètes.

Il est souvent commode d'enregistrer les programmes sur une bande magnétique dans un ordre adapté à la méthode d'essai. La longueur des silences entre les programmes dépend également de la méthode d'essai et devra avoir une durée suffisante pour permettre de noter l'évaluation d'essai. Cinq secondes à 15 s suffiront normalement pour l'évaluation, mais l'expérimentateur peut avoir besoin d'un intervalle plus long pour rebobiner la bande. La pause entre les programmes répétés dans une comparaison par paires devra être de 1 s environ et ne pas dépasser 2 s.

La bande de programme doit contenir également les signaux nécessaires à l'étalonnage de la réponse en fréquence et du niveau sonore.

2.5 Réglage de niveau

2.5.1 Niveaux d'écoute des programmes

Pour chaque programme, le niveau sonore reproduit devrait être, en principe, le niveau d'écoute préféré par l'auditeur moyen. L'expérience prouve que le niveau sonore préféré est souvent très proche du niveau de la parole ou de la musique originale écoutée dans la position normale de l'auditeur pendant l'exécution réelle. Les évaluations de fidélité (voir paragraphe 3.4 et annexe A) des sons reproduits seraient probablement également plus faciles si les niveaux d'écoute relatifs aux différents programmes reflétaient les niveaux relatifs des créations originales.

En variante, le niveau d'écoute moyen préférentiel peut être évalué par des essais d'écoute avec chaque programme, effectués dans le local d'écoute, en utilisant un haut-parleur de haute qualité.

Il est souvent souhaitable d'introduire plus d'un niveau d'écoute dans un essai, ceci pour deux raisons. Premièrement, il existe de grandes différences entre les niveaux d'écoute préférentiels des auditeurs, et certains auditeurs peuvent préférer des niveaux considérablement inférieurs aux niveaux moyens. Deuxièmement, certains petits haut-parleurs peuvent être surchargés par des niveaux qui s'adaptent mieux à des haut-parleurs plus grands. Si l'on utilise des niveaux d'écoute supplémentaires, ils devraient être choisis par paliers de 10 dB au-dessous du niveau initial. Toutefois, pour limiter la durée totale de l'essai, il n'est pas nécessaire, en général, de répéter tous les programmes à chaque niveau d'écoute. Par contre, seuls un ou deux programmes ayant le niveau le plus élevé peuvent être reproduits à des niveaux inférieurs de 10 dB ou 20 dB.

2.5.2 Niveaux relatifs des haut-parleurs

Pour éviter d'influencer les résultats d'essai, il est essentiel que chaque programme ait la même intensité sonore dans tous les haut-parleurs. En général, cela est obtenu par le réglage des commandes de gain dans le système de sélection des haut-parleurs, en utilisant la méthode de mesure suivante:

In the monophonic tests special care should be taken in the selection of programme material to ensure that audible colorations do not occur due to the combination of two stereophonic channels into one. Such a requirement will eliminate many commercial stereophonic recordings.

Note. — Because uncontrolled variations exist in commercial recordings (the number of microphones used, their type and placement, equalization, added reverberation, etc.) it is sometimes advisable to use musically redundant sections from a variety of sources (different recording companies, studios, engineers, etc.). The music and its sources should always be quoted in the test report.

The preparation of an internationally recognized standard programme tape is under consideration. Such a tape would facilitate the comparison of listening tests done at different places.

Where the musical programme is to be used to test other than high fidelity system components (e.g., a.m. receivers), band limiting is permitted. The precise frequency weighting function must be specified in these instances.

Programme sections should be of approximately equal durations, in the range 20 s to 40 s. For pair-wise comparisons identical sections are used in a pair. Sections of music should as far as possible form complete musical phrases.

It is often convenient to record the programme sections on magnetic tape in an order suited to the test procedure. The length of silent periods between sections will also depend on the test procedure and must be long enough to permit the writing down of the test score. Five seconds to 15 s is normally sufficient for the scoring, but a longer interval may be necessary for the experimenter to rewind the tape. The pause between repeated sections in a paired comparison should be about 1 s and not longer than 2 s.

The programme tape shall also contain necessary frequency-response and sound-level calibration signals.

2.5 *Level setting*

2.5.1 *Listening levels for programmes*

For each programme section, the reproduced sound level should, in principle, be the preferred listening level for the average listener. There is some experimental evidence that the preferred sound level is often fairly close to the level of the original speech or music at a typical listener's position during the live performance. The task of making fidelity ratings (see Sub-clause 3.4 and Appendix A) of reproduced sounds would probably also be more easily done if the relative listening levels for different programme sections reflected the relative levels of the original performances.

Alternatively, the preferred average listening level may be assessed by listening tests on each programme section, carried out in the listening room using a high quality loudspeaker.

It is often desirable to include more than one listening level in a test for two reasons. One reason is that there are large differences between the preferred listening levels of the individual listeners, and some listeners may prefer considerably lower levels than the average. Another reason is that some small loudspeakers may be overloaded by levels which are preferred for bigger loudspeakers. If additional listening levels are used, they should be chosen in steps of 10 dB below the primary level. However, in order to limit the total test length, it is not generally necessary to repeat all programme sections at each listening level. Instead, only those one or two programme sections having the highest level in the test might be reproduced at a level 10 dB or 20 dB lower.

2.5.2 *Relative levels for loudspeakers*

To avoid biasing the test results, it is essential that each individual programme section should sound equally loud over all of the loudspeakers. In general, this is achieved by adjusting the gain controls in the loudspeaker selection system using the following measurement procedure:

- a) Signal d'essai: Bruit permanent avec un spectre représentant les caractéristiques de la moyenne de tous les types de programme, conformément à la Publication 268-1 de la CEI: Equipements pour systèmes électroacoustiques, Première partie: Généralités.
- b) Appareils Sonomètre conforme à la Publication 651 de la CEI: Sonomètres, courbe de mesure: pondération A, classe 1, caractéristique temporelle S.

Le niveau sonore est mesuré dans le local d'écoute à la hauteur de l'oreille, au milieu des places assises, lorsque les haut-parleurs sont alimentés avec le signal d'essai et que le gain est réglé de telle sorte que la lecture du sonomètre soit identique pour tous les haut-parleurs.

Dans certains cas, en fonction du spectre du programme et de l'équilibre spectral des haut-parleurs, l'équilibrage du niveau sonore obtenu par les mesures n'est pas suffisant. Il est alors conseillé d'effectuer un équilibrage minutieux en appliquant des méthodes subjectives. Dans les situations vraiment critiques, il peut s'avérer nécessaire d'effectuer des équilibrages individuels du niveau sonore subjectif pour tous les programmes.

2.6 Conditions électriques

Si l'essai d'écoute doit être un élément de jugement valable des différences entre les haut-parleurs, il est essentiel de porter une attention particulière à certaines conditions électriques.

2.6.1 Amplificateur de puissance

L'amplificateur de puissance doit être capable d'alimenter tous les haut-parleurs d'essai au niveau sonore le plus élevé, sans écrêtage ni instabilité ni déclenchement des circuits de protection. C'est une bonne habitude que de surveiller la tension de sortie de l'amplificateur, du moins au cours des essais préliminaires, en utilisant un oscilloscope ou un indicateur de puissance de crête. La combinaison des impédances complexes de certains haut-parleurs avec celles des fils de raccords peuvent soulever des problèmes avec certains amplificateurs de puissance.

2.6.2 Système de commutation

Il est d'usage courant dans les essais d'écoute d'utiliser un système de commutation pour choisir les différents haut-parleurs en essai. Un tel système ne doit compromettre ni le fonctionnement des haut-parleurs ni celui de l'amplificateur de puissance. La résistance de contact du commutateur, en particulier dans le circuit de sortie de l'amplificateur, doit être très basse et il faut vérifier régulièrement sa valeur (voir paragraphe 2.6.3). Les réglages de niveau (voir paragraphe 2.5) doivent être réalisés au moyen de potentiomètres placés avant l'entrée de l'amplificateur de puissance.

Les connexions de masse d'entrée et de sortie de l'amplificateur de puissance doivent être séparées électriquement l'une de l'autre, en utilisant une commutation distincte entre l'entrée et la sortie. Une masse commune dans le système de commutation peut, avec certains amplificateurs, entraîner des modifications de réponse en fréquence en liaison avec la caractéristique d'impédance des haut-parleurs.

On vérifiera par des mesures que l'interconnexion de l'appareil n'a pas introduit de changements sensibles dans des paramètres de qualité importants, tels que réponse en fréquence ou bruit de fond.

2.6.3 Fils de raccordement

Il convient que les fils reliant la sortie de l'amplificateur de puissance aux haut-parleurs soient aussi courts que possible et, en tout cas, d'une section suffisante pour que la résistance totale de n'importe quel circuit de haut-parleur (mesurée à partir de la sortie de l'amplificateur de puissance, à travers tous les contacts des commutateurs jusqu'aux bornes court-circuitées des haut-parleurs) ne soit supérieure ni à 0,2 Ω ni à 1/40 de l'impédance nominale la plus basse des haut-parleurs en essai, la plus petite des deux valeurs étant applicable.

- a) Test signal: Continuous noise with a spectrum representing properties of the average of all types of programme material according to IEC Publication 268-1: Sound System Equipment. Part 1: General.
- b) Measuring instrument: Sound level meter according to IEC Publication 651: Sound Level Meters, weighting curve A, Class 1, time characteristic S.

The sound level is measured in the listening room at ear level in the middle of the seating area when the loudspeakers are fed with the test signal, and the gain is adjusted until the meter reads the same for all loudspeakers.

Occasions may arise, depending on the spectrum of the programme and the spectral balance of the loudspeakers, when the loudness balances achieved by measurement are not adequate. In these cases, careful balancing by subjective methods is advised. In very critical situations, it may be necessary to do individual subjective loudness balances for every programme section.

2.6 *Electrical requirements*

If the listening test is to be a valid indicator of differences between loudspeakers it is essential that careful consideration be given to certain electrical requirements.

2.6.1 *Power amplifier*

The power amplifier shall be capable of driving all test loudspeakers at the highest sound level without clipping, instability or triggering of protective circuitry. It is good practice to monitor the amplifier output, during at least preliminary tests, using an oscilloscope or peak power indicator. The complex impedance of some loudspeaker and connecting wire combinations can cause problems with certain power amplifiers.

2.6.2 *Switching system*

It is commonplace in listening tests to use a switching system to select the different loudspeakers under test. Such a system shall not compromise the performance of either the loudspeakers or the power amplifier. Switch contact resistance, especially in the amplifier output circuit, shall be very low and shall be checked periodically for deterioration (see Sub-clause 2.6.3). Level settings (see Sub-clause 2.5) shall be accomplished by means of potentiometers positioned before the input to the power amplifier.

Power amplifier input and output earth connections shall be electrically isolated from each other by using separate input and output switching. A common earth connection through the switching system can, with some amplifiers, result in frequency-response modifications related to loudspeaker impedance characteristics.

It should be confirmed by measurement that the interconnection of the apparatus has not introduced significant changes in important performance parameters, such as frequency-response or background noise.

2.6.3 *Connecting wires*

Wires from the power amplifier output to the loudspeakers should be as short as possible and in any event shall be of sufficient gauge that the total resistance of any loudspeaker circuit (measured from the power amplifier output, through all selector switch contacts to the short-circuited loudspeaker terminals) does not exceed $0.2\ \Omega$ or $1/40$ of the lowest rated impedance of the loudspeakers under test, whichever is less.

2.6.4 Réponse en fréquence

Certains haut-parleurs sont destinés à être utilisés avec une compensation de réponse en fréquence sur le signal électrique. Une telle prescription peut être satisfaite en introduisant un dispositif spécial dans le système, ou en réglant la réponse en fréquence sur le signal électrique, au moyen d'un correcteur courant, pour suivre la recommandation du constructeur. Sauf instruction spécifique contraire, les réglages de réponse amplitude/fréquence devront être effectués avant les réglages des niveaux de puissance.

Dans d'autres cas, par exemple pour des haut-parleurs bon marché, il peut être courant pour les utilisateurs d'appliquer la compensation de réponse en fréquence par le réglage normal de la tonalité ou par le correcteur physiologique de l'amplificateur associé. Pour que les résultats de l'essai d'écoute soient réalistes, il peut être nécessaire d'ajouter des essais simulant cette pratique aux essais normalisés plus rigoureux.

Tout essai dans lequel la réponse en fréquence du quadripôle électrique n'est pas droite doit être décrit dans le rapport joint aux résultats d'essai avec un relevé de la réponse en fréquence.

3. Procédure expérimentale et évaluation

3.1 Auditeurs

Le choix des auditeurs dépend de la population des auditeurs, pour lesquels les résultats d'essai sont supposés être valables. Il est bien connu que des groupes d'auditeurs ayant différentes habitudes d'écoute «haute fidélité», des formations musicales différentes, etc., ont des appréciations différentes. Dans le cadre d'un groupe ayant des expériences similaires, la fiabilité inter-individuelle est normalement très grande et, en conséquence, le nombre d'auditeurs dans un tel groupe peut être plutôt restreint (toutefois, quatre auditeurs doivent être considérés comme un minimum, voir annexe C).

D'autre part, le nombre de groupes dépend de la généralité souhaitée pour l'essai. Pour les essais de recherche et de développement, un groupe d'auditeurs avertis peut être suffisant (ayant l'expérience de l'écoute «haute fidélité» et aussi, de préférence, de la musique non enregistrée). Si les résultats doivent être généralisés pour une population spécifiée d'auditeurs, les participants devront être sélectionnés de façon à être certains qu'ils représentent la population concernée (voir également l'annexe C).

La fiabilité des évaluations devra être vérifiée pendant le traitement statistique des résultats (voir annexe C) et devra être considérée comme une partie importante des informations contenues dans les résultats d'essai. Si la fiabilité n'est pas considérée comme satisfaisante, il est préférable d'introduire davantage d'auditeurs dans l'essai.

On devra vérifier l'audition normale des auditeurs (niveaux de seuil d'audition inférieurs à 20 dB h.l. (Norme ISO 389) dans la gamme de fréquences de 125 Hz à 8 000 Hz). Les personnes ayant des troubles d'audition ne doivent pas être forcément exclues en tant qu'auditeurs, mais leurs résultats d'essai devront être analysés séparément pour voir s'il est raisonnable de mettre en commun leurs observations avec celles des auditeurs à audition normale.

3.2 Durée de l'essai

Pour des raisons pratiques, la durée d'une seule séance d'essai devra être limitée à 1 h environ, avec une ou deux pauses. La durée totale dépend de facteurs tels que la méthode d'essai, le nombre d'appareils mis à l'essai, le nombre souhaité d'auditeurs et de répétitions présentées à chacun d'eux. Des exemples de la façon d'évaluer la durée de l'essai sont donnés dans l'annexe B. Pour des essais d'intelligibilité, il peut être nécessaire de soumettre chaque auditeur à plus d'un essai.

Un auditeur ne doit pas dépasser 2 h d'écoute par jour.

2.6.4 Frequency response

Certain loudspeakers are designed to be used with frequency-response compensation in the electrical signal path. Such a requirement may be met by the insertion of a dedicated device into the system or by an adjustment of the signal-path frequency response, using a conventional equalizer, to correspond to the manufacturer's recommendation. In the absence of specific instructions to the contrary, the amplitude/frequency response adjustments should be made prior to the power level adjustments.

In other cases, such as with inexpensive loudspeakers, it may be a common practice for users to apply frequency response compensation through conventional tone controls or loudness-compensating devices in the associated amplifying device. For the results of the listening test to be realistic it may be necessary to include tests simulating this practice, in addition to the more demanding standardized tests.

Any tests in which the frequency response of the electrical signal path is not flat shall be described in the report accompanying the test results and the frequency response shown.

3. Experimental procedure and evaluation

3.1 Listeners

The choice of listeners depends on the population of listeners for which the test results are supposed to be valid. It is known that there are differences in the judgements of listener groups having different experience of "high-fidelity" listening, musical training, etc. Within a group of similar experience the inter-individual reliability is normally quite high and therefore the number of listeners in such a group can be rather small (however, four listeners should be considered as a minimum, see Appendix C).

The number of groups, on the other hand, depends on the desired generality of the test. For tests in research and development one group of experienced listeners may be sufficient (experienced in "hi-fi" listening and preferably also in listening to live music). If the results are to be generalized for a specified population of people, the listeners should be selected in such a way as to ensure that they are representative of the population in question (see also Appendix C).

The reliability of the ratings should be checked during the statistical treatment of the data (see Appendix C) and should be considered as an important part of the information contained in the test results. If the reliability is judged not to be satisfactory it may be preferable to include more listeners in the test.

The listeners should be checked for normal hearing (hearing threshold levels less than 20 dB h.l. (ISO Standard 389) in the frequency range of 125 Hz to 8 000 Hz). Hearing-impaired people need not necessarily be excluded as listeners, but their test results should be analyzed separately to see whether it is reasonable or not to pool their data with data from normal-hearing listeners.

3.2 Test duration

For practical reasons the length of a single-test session should be limited to about 1 h, with one or two breaks. The total duration depends on such factors as the test procedure, the number of systems included in the test, the desired number of listeners and replications presented to each listener. Examples of how the test duration can be estimated are given in Appendix B. For comprehensive listening tests it may be necessary to use more than one test session for each listener.

No listener should be engaged in more than 2 h of listening per day.

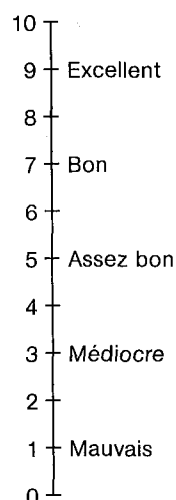
3.3 Echelles d'évaluation

La qualité sonore globale doit être évaluée suivant une échelle de fidélité. On demande aux auditeurs de juger dans quelle mesure le son reproduit ressemble au son idéal. Toutefois, la grande variété des groupes musicaux et méthodes d'enregistrement impose d'utiliser des critères différents pour apprécier la reproduction d'enregistrements différents.

- a) Pour la musique exécutée avec des instruments classiques et pour le chant et la parole produits par des voix naturelles, l'auditeur peut faire appel à sa mémoire pour se rappeler des expériences analogues réelles et pour s'en servir comme référence pour juger dans quelle mesure la reproduction correspond au son réel original. Un tel critère de «conformité à la nature» s'applique typiquement à de nombreux enregistrements de musique classique, de musique folklorique, de musique de jazz, etc.
- b) D'autre part, un nombre élevé et toujours croissant d'enregistrements contemporains de tous les types de musique ne cherche pas à recréer pour les auditeurs une reproduction rigoureusement exacte d'une exécution réelle. Le choix et l'emplacement des microphones, le mixage de nombreuses entrées distinctes, l'addition de retards produits électroniquement, la réverbération et la correction de la réponse en fréquence, sont tous guidés par un goût artistique basé sur les sons reproduits par les haut-parleurs de contrôle dans une cabine d'enregistrement, c'est-à-dire dans une ambiance d'écoute qui est, en général, très différente des conditions domestiques typiques et, naturellement, des recommandations données dans le présent rapport. En conséquence, dans ces cas et dans le cas d'une musique synthétisée électroniquement, les auditeurs ne peuvent que rapporter l'essai d'écoute à un idéal imaginé, à un concept de ce qui était recherché.

On doit tenir compte de telles considérations dans le choix du programme, comme indiqué dans la deuxième note du paragraphe 2.4. Il peut également être utile de suggérer aux auditeurs à quel critère de jugement ils devront s'attacher, en évaluant certaines séquences individuelles de programme.

Dans tous les cas, il est demandé aux auditeurs de rapporter leurs jugements de fidélité à une échelle de 0 à 10, où 0 représente la pire reproduction imaginable et où 10 représente ce que l'auditeur croit être l'objectivité parfaite de l'enregistrement. On peut s'attendre à ce que les notes de 1 à 9 sur cette échelle graduée soient réellement utilisées, alors que les valeurs de 0 et 10 sont considérées comme des bornes définissant les extrémités de l'échelle. Le long de l'échelle, les définitions verbales «Excellent — Bon — Assez bon — Médiocre — Mauvais» (ou leurs traductions dans d'autres langues) aideront les auditeurs à évaluer la qualité des haut-parleurs. Une description plus détaillée de l'échelle d'évaluation, présentée ci-dessous, figure dans l'annexe A.



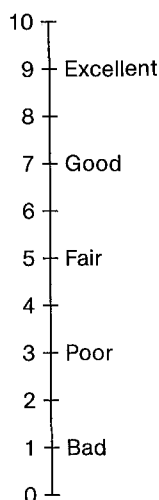
3.3 Rating scales

The overall sound quality is to be rated on a scale of fidelity. Listeners are required to judge the extent to which the reproduced sound resembles an ideal. However, the wide variety of music and recording methods makes necessary the use of different criteria in judging the reproduction of different recordings.

- a) For music with conventional acoustic instruments and for singing and speech with natural voices, the listener can use his memory of similar live experiences as a reference and judge how well the reproduction matches the original live sound. Such a "true-to-nature" criterion would apply, typically, to many recordings of classical music, folk music, jazz etc.
- b) On the other hand, a large and ever-increasing number of contemporary recordings of all types of music do not attempt to recreate for listeners a strictly accurate reproduction of a live performance. The choice and placing of microphones, the mixing of the many separate inputs, the addition of electronically generated delays, reverberation and frequency-response equalization are all guided by an artistic judgement based on sounds reproduced over monitor loudspeakers in a recording studio control room, a listening environment which generally will be very different from typical household conditions and, of course, the room recommendations given in this report. Consequently, in these cases and in the case of electronically synthesized music, listeners can only relate the test reproduction to an imagined ideal, a concept of what was intended.

Such considerations should be taken account of in the choice of programme material, as mentioned in the second note to Sub-clause 2.4. It may also be necessary to suggest to the listeners which judgement criterion they should apply when scoring the individual programme sections.

In all cases listeners are required to report their fidelity judgements on a scale of 0 to 10, where 0 represents the worst imaginable reproduction and 10 represents what the listener imagines to be the objective of the recording. On this interval scale it is expected that ratings from 1 to 9 will actually be used, while scale values of 0 and 10 are regarded as "anchors" defining the end points of the scale. Along the scale the verbal descriptions "Excellent — Good — Fair — Poor — Bad" (or their translations into other languages) will assist listeners in scoring the performance of the loudspeakers. A more detailed description of the rating scale, shown below, appears in Appendix A.



La qualité sonore globale peut être considérée comme une qualité multidimensionnelle, c'est-à-dire qu'elle est constituée par une certaine combinaison de dimensions sensorielles distinctes. Par conséquent, il peut être utile de compléter ces évaluations de la qualité globale par des évaluations de qualités spécifiques. On ne dispose pas, actuellement, de connaissances suffisantes pour justifier des recommandations précises concernant les types et la qualité d'échelles complémentaires à inclure. En attendant des résultats ultérieurs sur ce sujet, la sélection des échelles complémentaires est donc laissée à l'initiative de l'utilisateur.

Note. — Une récapitulation de la recherche actuelle propose les dimensions sensorielles suivantes, comme composantes de la qualité sonore globale :

Clarté, netteté, définition
Agressivité, dureté
Brillance
Plénitude
Ampleur, franchise
Distance
Bruits, sons perturbateurs
Intensité sonore

L'importance relative de ces dimensions est probablement différente dans des contextes différents et pour des auditeurs différents. On s'attend que la recherche future procure des informations plus détaillées qui pourront être utilisées dans les essais d'écoute pratiques.

3.4 Choix des méthodes d'essai

Une règle fondamentale d'un essai d'écoute est qu'il soit conçu comme une expérience dirigée. La variable dépendante de l'essai est l'évaluation de l'auditeur. Les variables indépendantes pouvant être manipulées par l'expérimentateur sont premièrement les haut-parleurs en essai et deuxièmement les séquences de programme. Les variables étrangères devront être éliminées ou rigoureusement contrôlées afin d'éviter qu'elles influencent les évaluations.

Deux méthodes d'essai différentes sont recommandées :

- i) Evaluations par stimuli isolés (S)
- ii) Comparaisons par paires (P)

La méthode d'évaluation par stimuli isolés implique que chaque «stimulus» soit présenté séparément et que l'évaluation soit faite par l'auditeur avant la présentation du stimulus suivant. Un stimulus est un programme donné (1, 2, 3, 4, etc.) reproduit par un haut-parleur donné (A, B, C, D, etc.).

D'autre part, la méthode des comparaisons par paires consiste à présenter tous les stimuli par paires, chaque paire comportant un certain programme reproduit d'abord par un haut-parleur et, immédiatement après, par un autre. Les diagrammes des temps pour la présentation du stimulus selon les deux méthodes figurent à l'annexe B.

L'échelle d'évaluation recommandée au paragraphe 3.3 s'applique aux deux méthodes.

Le choix de la méthode à appliquer, S ou P, dépend du nombre de haut-parleurs, du nombre souhaité de répétitions (c'est-à-dire d'écoutes répétées de chaque stimulus pour augmenter la fiabilité des évaluations) et du temps disponible pour effectuer l'essai. Une comparaison entre S et P, tenant compte de ces facteurs, est indiquée dans l'annexe B.

On peut constater que pour un nombre allant jusqu'à six haut-parleurs inclus dans le même essai, la méthode P ne conduit pas, en général, à des durées d'essai sensiblement plus longues que celles de la méthode S, du fait de la pause plus courte entre les présentations par paires et également parce que certaines ou toutes les répétitions souhaitées sont obtenues par le nombre nécessaire de comparaisons par paires (voir paragraphe 3.6). La méthode P est préférée par la plupart des

The overall sound quality can be considered as a multidimensional attribute, that is, it is constituted by some combination of separate perceptual dimensions. It may therefore be useful to supplement the judgements on overall quality with judgements on specific perceptual dimensions. At present there is insufficient knowledge available to justify definite recommendations about which supplementary scales, and how many, should be included. Awaiting further results on this topic, the selection of supplementary scales is therefore left to the user's consideration.

Note. — A survey of present research suggests the following perceptual dimensions as constituents of the overall sound quality:

Clarity, distinctness, definition
Sharpness, hardness
Brightness
Fullness
Spaciousness, openness
Distance
Noises, interfering sounds
Loudness

The relative importance of these dimensions is probably different in different contexts and for different listeners. It is expected that future research will provide more detailed information that can be used in practical listening tests.

3.4 *Choice of test procedure*

A basic requirement of a listening test is that it should be designed as a controlled experiment. The dependent variable in the test is the listener's judgement. The independent variables to be manipulated by the investigator are primarily the loudspeakers under test and secondarily the programme sections. Extraneous variables should be eliminated or held under due control to avoid biasing the judgements.

Two different test procedures are recommended:

- i) Single stimulus ratings (S)
- ii) Paired comparisons (P)

The method of single stimulus rating implies that each "stimulus" is presented separately and a rating is made by the listener before the next stimulus is presented. A stimulus comprises a given programme (1, 2, 3, 4, etc.) reproduced by a given loudspeaker (A, B, C, D, etc.).

The method of paired comparisons, on the other hand, means that all stimuli are presented in paired sequences, each pair consisting of a certain programme section reproduced first over one loudspeaker and immediately afterwards over another loudspeaker. The time patterns for the stimulus presentation in both procedures are illustrated in Appendix B.

The rating scale as recommended under Sub-clause 3.3 is applicable for both methods.

The choice of procedure, S or P, that should be applied depends on the number of loudspeakers, the desired number of replications (that is, repeated applications of each stimulus in order to increase the reliability of the ratings) and the time available for doing the test. A comparison between S and P taking these factors into account is given in Appendix B.

It may be noted that for up to six loudspeakers included in the same test the P procedure does not in general lead to noticeably longer test times than the S procedure because of the shorter pause between the presentations in a pair and also because some or all of the desired replications are obtained by the necessary number of paired comparisons (see Sub-clause 3.6). The P procedure is preferred by most listeners and in particular when several of the systems under test

auditeurs et, en particulier, lorsque plusieurs des systèmes en essai sont de qualité à peu près égale, de sorte qu'il faille faire des distinctions plus subtiles. Il est possible de combiner les deux méthodes, en commençant par S pour un vaste groupe de haut-parleurs et ensuite en appliquant P pour les systèmes qui ont été évalués comme étant à peu près de même niveau.

Notes 1. — Quelle que soit la méthode d'essai, un ou plusieurs haut-parleurs de référence peuvent être inclus dans certains ou tous les essais. Cela facilite la comparaison des résultats entre les différents essais.

2. — Une modification de la méthode P recommandée consiste à s'abstenir de faire des évaluations numériques, en demandant simplement aux auditeurs de préciser quelle interprétation (haut-parleur) dans la paire ils préfèrent en ce qui concerne la fidélité de la reproduction. Les résultats obtenus ne seront donc pas des évaluations numériques directes des haut-parleurs, mais indiqueront seulement le nombre d'essais dans lesquels chaque haut-parleur est préféré à un autre. Cela peut être utilisé pour une classification des haut-parleurs. Si on le désire, cette hiérarchie peut être transformée en une échelle d'évaluation par l'utilisation de modèles comme «la loi des jugements comparatifs» ou de modèles pour la graduation «non métrique». La méthode simplifiée préférentielle donne couramment moins de renseignements que la méthode P recommandée en ce qui concerne l'importance des différences entre les haut-parleurs. La procédure recommandée donne également plus de liberté dans le traitement des résultats en ce sens que les évaluations numériques sont traitées directement comme spécifié au paragraphe 3.7 et dans l'annexe C, ou bien qu'elles peuvent être également utilisées pour déterminer, par le calcul, le nombre d'essais, dans lesquels chaque haut-parleur est préféré à un autre, comme l'explique cette note.

Pour contrebalancer l'influence sur les évaluations de l'habitude, de la fatigue ou d'autres variations liées au temps, il convient d'utiliser un ordre aléatoire de présentation des stimuli (c'est-à-dire combinaisons de programmes et de haut-parleurs). Un ordre différent doit être utilisé pour chaque séance d'essai.

Dans le cas de la méthode P, la règle de l'ordre aléatoire de présentation s'applique à l'ordre des paires de stimuli (voir figure de l'annexe B). Dans le cadre de chacune de ces paires, l'ordre de présentation de deux haut-parleurs associés doit être équilibré (c'est-à-dire que l'ordre A-B et B-A devra se produire le même nombre de fois dans chaque programme).

Si des raisons pratiques l'exigent, une sorte d'ordre de présentation systématique peut être utilisée au lieu d'un ordre entièrement aléatoire, à condition que les effets cumulatifs éventuels (dus à des facteurs tels que la pratique, l'adaptation, la fatigue, etc.) soient répartis de façon égale parmi les conditions des stimuli dans l'essai. Toutefois, il convient de remarquer que l'application de l'analyse de variance indiquée dans l'annexe C suppose a priori un ordre de présentation entièrement aléatoire.

La mise en œuvre réelle d'un essai d'écoute conforme à l'une ou à l'autre des deux méthodes recommandées doit être facile à comprendre d'après les annexes A et B.

3.5 Instructions et essais préliminaires

Les instructions données aux auditeurs dans un essai d'écoute ont un effet sensible sur leur comportement et leur jugement. Le point décisif est la définition de l'échelle d'évaluation ainsi que les indications concernant la manière de son utilisation. C'est pourquoi, un libellé particulier de cette partie de l'instruction doit être joint à chaque méthode d'essai normalisé et utilisé à la lettre; voir l'annexe A.

Il est souhaitable d'introduire quelques observations sur la possibilité qu'ont les auditeurs de savoir quel était le son original réel de la musique (ou parole) ou comment ce son était censé être. L'importance de ces observations dépend, toutefois, des programmes et des auditeurs choisis pour l'essai.

Les instructions devraient être données à la fois verbalement (de préférence sur une bande enregistrée) et par écrit.

are of about equal quality, so that finer discriminations have to be made. A combination of both procedures is possible, starting with S for a large group of loudspeakers and then proceeding with P solely on those systems which have received about the same rating.

Notes 1. — Independently of the test procedure, one or more reference loudspeakers may be included in some or all tests. This facilitates the comparison of results between different tests.

2. — A modification of the recommended P procedure is to refrain from numerical judgements, asking the listener(s) merely to state which reproduction (loudspeaker) in the pair they prefer with regard to the fidelity of the reproduction. Thus the resulting data will not be direct numerical judgements of the loudspeakers, but only give the number of tests in which each loudspeaker is preferred to another loudspeaker. This can be used for a ranking of the loudspeakers. If wanted, these rank order data may possibly be transformed into an interval scale by use of models like "the law of comparative judgements" or models for so-called non-metric scaling. The simplified preference procedure will usually give less information than the recommended P procedure with regard to the size of the differences between the loudspeakers. The recommended procedure also gives more freedom in the data treatment in the sense that either the numerical judgements are treated directly as described in Sub-clause 3.7 and Appendix C, or they can also be used to compute the number of tests in which each loudspeaker is preferred to another loudspeaker as discussed in this note.

To counteract the influence on the judgements due to practice, fatigue, or other variations with time, a random presentation order of the stimuli (i.e. combinations of programmes and loudspeakers) should be used. A different order should be used for each test session run.

In the P procedure the requirement for random presentation order refers to the order of the stimulus pairs (see figure in Appendix B). Within each such pair the presentation order of the respective two loudspeakers should be balanced (that is, the order A-B and B-A should occur equally often for each programme section).

If necessary for practical reasons, some kind of systematic presentation order may be used instead of a completely random one, provided that the possible progressive effects (due to such factors as practice, adaptation, fatigue, etc.) are distributed equally over the stimulus conditions in the test. However, it should be noted that the application of analysis of variance as outlined in Appendix C presupposes completely random order of presentation.

The actual implementation of a listening test according to either of the two recommended procedures should be clear from Appendices A and B.

3.5 *Instruction and preliminary trials*

The instructions given to the listeners in a listening test have a noticeable effect on their behaviour and judgement. The crucial part is the definition of the judgement scale as well as indications of how it is to be used. Therefore, a specific wording of this part of the instruction should be included with each standardized test procedure and should be used word for word; see Appendix A.

It is desirable to include some comment upon the listeners' possibility of knowing how the music (or speech) actually sounded in its original form or how it is intended to sound. The importance of this will, however, depend on the programme sections and listeners chosen for the test.

The instruction should be given both orally (preferably from a tape-recording) and in written form at the same time.

Les instructions devront être suivies d'un certain nombre d'essais préliminaires destinés à faciliter l'adaptation aux conditions de jugement. Dans ces essais, les stimuli devraient être représentatifs de ceux de l'essai réel. Si l'essai complet dure environ 1 h, la séance d'entraînement doit durer au moins 7 min à 8 min. Pour les séances d'essai permanentes, le nombre d'essais préliminaires peut être réduit progressivement. Il est également recommandé de faire quelques essais de « mise en train » aussitôt après la pause en milieu de séance.

3.6 *Nombre de jugements*

Une condition minimale est que chaque auditeur puisse juger chaque stimulus (= reproduction d'un programme donné par un haut-parleur donné) au moins une fois. Toutefois, deux répétitions pour chaque stimulus sont nécessaires, et trois ou quatre sont souvent souhaitables, pour permettre le calcul de la fiabilité de jugement d'un auditeur.

Dans le cas de la méthode d'évaluation par stimuli isolés, le nombre de répétitions peut être choisi en même temps que la durée totale de l'essai et le nombre d'auditeurs de façon à obtenir la fiabilité requise.

Dans le cas de la méthode P utilisant n haut-parleurs, le nombre de répétitions pour chaque stimulus sera automatiquement $n - 1$, étant donné que chaque haut-parleur forme une paire avec chacun des autres $n - 1$ haut-parleurs pour chaque programme. (Le nombre total de présentations des haut-parleurs par paires pour chaque programme est $n(n - 1)/2$.) Si le nombre de haut-parleurs est suffisamment grand, la fiabilité requise sera probablement atteinte sans autres répétitions.

Pour les deux méthodes, voir les annexes B et C pour de plus amples détails.

3.7 *Traitement statistique des résultats*

Une description de l'analyse statistique est donnée dans l'annexe C. On ne donne ici que les principes essentiels.

Les résultats de jugement des auditeurs sont introduits dans des tableaux appropriés (programme $\times$ haut-parleur) pour chaque participant séparément et/ou pour tous les auditeurs ensemble. Les moyennes arithmétiques sont calculées dans toutes les combinaisons (programme $\times$ haut-parleur), pour chaque haut-parleur en faisant la moyenne des notes correspondant aux divers programmes, pour les programmes en faisant la moyenne des notes obtenues avec les divers haut-parleurs. On procède à l'examen visuel de ces tableaux et éventuellement des présentations graphiques correspondantes pour arriver à des conclusions sur l'évaluation des qualités des haut-parleurs en général et en combinaison avec divers programmes.

Si on a besoin d'une analyse statistique plus détaillée, on peut l'obtenir au moyen d'une analyse de variance et d'autres essais statistiques associés. Cela permet d'évaluer l'interdépendance de jugement entre les haut-parleurs, les programmes, les auditeurs et les diverses interactions entre ces facteurs. Les analyses de variance peuvent être utilisées également pour estimer la fiabilité des résultats de jugement, aussi bien pour chaque auditeur séparément (fiabilité individuelle) que pour tous les auditeurs ensemble (fiabilité interindividuelle). Si des variables supplémentaires sont introduites dans l'essai, par exemple positions des haut-parleurs ou des auditeurs, l'analyse statistique décrite peut être généralisée pour couvrir également ces variables (annexe C, article C5).

3.8 *Contenu d'un rapport d'essai*

Le rapport d'un essai d'écoute devra comporter, en principe, des informations appropriées à toutes les questions traitées dans les articles 2 et 3 du présent rapport.

After the presentation of the instruction a number of preliminary trials should be made to facilitate adaptation to the judgement situation. The stimuli in these trials should be representative of those in the real test. If the duration of the complete test is about 1 h, the training session should last at least 7 min to 8 min. For continued test sessions the number of preliminary trials may be successively decreased. It is also recommended to have a few "warming-up" trials directly after the break in the middle of a session.

3.6 *Number of judgements*

A minimum requirement is that each listener shall judge each stimulus (= reproduction of a certain programme section over a certain loudspeaker) at least once. To make possible a calculation of the reliability of a listener's judgement, however, two replications for every stimulus are necessary, and three or four replications are often desirable.

With the single stimulus rating procedure, the number of replications may be chosen together with the total test length and the number of listeners so as to yield the desired reliability.

In the P procedure with n loudspeakers the number of replications for each stimulus will automatically be $n - 1$, since each loudspeaker is paired with each of the other $n - 1$ loudspeakers for each of the programme sections. (The total number of paired presentations of loudspeakers for each programme is $n(n - 1)/2$.) If the number of loudspeakers is sufficiently large, the desired reliability will probably be reached without still further replications.

For both methods, see Appendices B and C for further details.

3.7 *Statistical treatment of the data*

A description of the statistical analysis is given in Appendix C. Only the main principles are given here.

The listeners' judgement data are entered into appropriate (programme $\times$ loudspeaker) matrices for each single listener and/or for all listeners together. Arithmetic means are computed for all (programme $\times$ loudspeaker) combinations, for loudspeakers in average over the programmes, and for programmes in average over the loudspeakers. Visual inspection of these matrices and of possible corresponding graphical displays is made to reach conclusions about the judged properties of the loudspeakers in general and in combination with various programme sections.

If a more detailed statistical analysis is required, this can be done by means of analysis of variance and other related statistical tests. This makes it possible to evaluate the dependency of the judgements on the loudspeakers, the programmes, the listeners, and on various interactions between these factors. The analyses of variance may also be used to estimate the reliability of the judgement data, both within each single listener (intra-individual reliability) and over all subjects together (inter-individual reliability). If extra variables are included in the test, for example, positions of loudspeakers or listeners, the described statistical analysis can be generalized to cover these variables too (Appendix C, Clause C5).

3.8 *Contents of test report*

The report of a listening test should in principle include relevant information on all items treated in Clauses 2 and 3 of this report.

En particulier, des spécifications précises devront être données sur les points suivants:

- a)* le local d'écoute et ses caractéristiques (dimensions, temps de réverbération, conditions atmosphériques, etc.),
- b)* les haut-parleurs et leur emplacement dans le local,
- c)* l'équipement électrique, d'éventuels écarts de réponse en fréquence,
- d)* le programme (descriptions et références aux disques ou bandes utilisés) et niveaux sonores utilisés pour la présentation,
- e)* le nombre et le type d'auditeurs et leur position dans le local,
- f)* la méthode d'essai détaillée (présentation du stimulus, échelles d'évaluation, instructions, nombre d'évaluations, etc.),
- g)* le traitement statistique des résultats (voir annexe C).

In particular, accurate specifications should be given for the following items:

- a)* the listening room and its characteristics (dimensions, reverberation time, atmospheric conditions etc.),
- b)* the loudspeakers and their position in the room,
- c)* the electrical equipment, and any frequency response deviations,
- d)* the programme material (descriptions and references to available gramophone records or tapes) and the sound levels used for presentation,
- e)* the number and type of listeners and their position in the room,
- f)* the test procedure in detail (stimulus presentation, judgement scales, instruction, number of judgements etc.),
- g)* the statistical treatment of the data (see Appendix C).

ANNEXE A

INSTRUCTIONS POUR LES ÉSSAIS D'ÉCOUTE

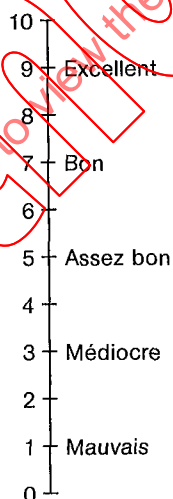
Ce qui suit peut servir de base aux instructions destinées aux auditeurs participant aux expériences soit par stimulus isolé, soit par la méthode de comparaison par paires. Il est important que tous les auditeurs reçoivent une préparation identique pour des essais de ce genre.

A1. Évaluations par stimuli isolés

«Au cours de cette expérience, vous entendrez de courts programmes de musique et de parole reproduits par un certain nombre de haut-parleurs. Tous les programmes seront présentés par tous les haut-parleurs dans un ordre aléatoire. Chaque fois que vous entendrez un programme, vous devrez juger la reproduction sur sa fidélité à un idéal. En ce qui concerne certaines musiques et paroles, vous pouvez faire appel au souvenir de vos expériences réelles pour évaluer dans quelle mesure la reproduction correspond au son réel original. Pour d'autres sortes de musique, par exemple musique légère ou électronique, vous devrez imaginer ce que l'on voulait vous faire entendre et évaluer dans quelle mesure la reproduction correspond au son voulu».

Note. — Pour éviter toute confusion éventuelle sur celui des critères ci-dessus qu'il fallait utiliser, l'expérimentateur peut suggérer aux auditeurs celui des programmes reproduits qu'il convient d'évaluer en termes de fidélité au son réel original, et celui qu'il convient d'évaluer en termes de fidélité au son «voulu».

«Dans tous les cas, les évaluations de la fidélité sont faites au moyen d'une échelle graduée de 0 à 10 comme indiqué ci-dessous:



Le nombre 10 indique une reproduction qui est parfaitement fidèle à l'idéal. Il n'existe pas d'amélioration possible.

Le nombre 0 indique une reproduction qui n'a aucune ressemblance avec l'idéal. Il est impossible d'imaginer une reproduction plus mauvaise.

«Tout nombre compris entre 0 et 10, comportant une décimale, peut être utilisé pour déterminer le degré de fidélité des reproductions. Écoutez attentivement tout le programme avant d'exprimer votre jugement. Enfin, tâchez de ne jamais oublier que vos préférences musicales ne doivent pas entrer en ligne de compte dans les évaluations qui se rapportent uniquement à la fidélité de la reproduction du son.»

Note. — Il convient que les instructions donnent également des informations sur la durée approximative de l'essai, les pauses, les essais préliminaires et d'autres détails utiles aux auditeurs.

APPENDIX A

INSTRUCTIONS FOR LISTENING TESTS

The following may be used as the basis for instructions to listeners involved in experiments of either single stimulus or paired-comparison formats. It is important that all listeners receive identical preparation for tests of these kinds.

A1. Single stimulus ratings

"In this experiment you will hear short programmes of music and speech reproduced over a number of loudspeakers. All of the programmes will be presented over all of the loudspeakers in a random sequence. Each time a programme is heard you should judge the reproduction in terms of its fidelity to an ideal. For some music and speech, you can use your memory of live experiences to rate how well the reproduction matches the original live sound. For other types of music, for example certain popular music and electronic music, you will have to imagine what was intended for you to hear and rate how well the reproduction matches the intended sound".

Note. — To avoid possible confusion about which of the above criteria should be used, the experimenter may suggest to the listener(s) which of the reproduced programmes should be rated in terms of fidelity to the original live sound, and which should be rated in terms of fidelity to the "intended" sound.

"In all cases the fidelity ratings shall be made on a scale from 0 to 10, as shown below:



The number 10 denotes a reproduction that is perfectly faithful to the ideal. No improvement is possible.

The number 0 denotes a reproduction that has no similarity to the ideal. A worse reproduction cannot be imagined.

"Any number between 0 and 10, including one decimal place, can be used to characterize the degree of fidelity in the reproductions. Listen carefully to the entire programme section before making your judgement. Finally, try to remember always that your musical preferences should not enter into the ratings, which refer only to the fidelity of the sound reproduction."

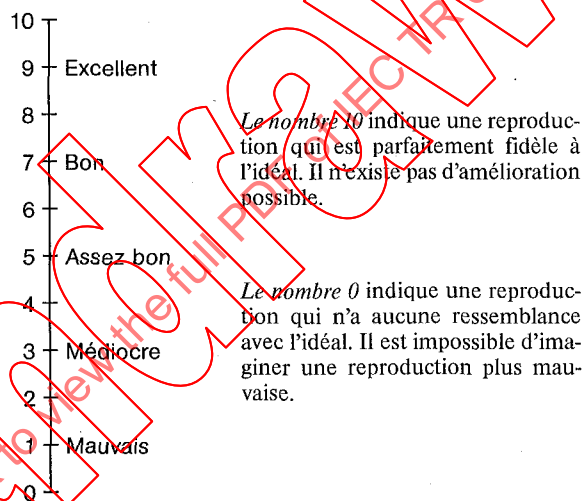
Note. — The instruction should also give information about the approximate test duration, pauses, preliminary trials and other details helpful to the listeners.

A2. Comparaisons par paires

«Au cours de cette expérience, vous entendrez de courts programmes de musique et de parole reproduits par un certain nombre de haut-parleurs. Chaque programme sera présenté deux fois, coup sur coup, chaque fois par un haut-parleur différent. L'ordre des programmes et la combinaison des haut-parleurs en paires varient d'une manière aléatoire. A la fin de chaque présentation par paires, vous devrez juger la reproduction par chacun des deux haut-parleurs du point de vue de sa fidélité à un idéal. Pour certaines sortes de musique et de parole, vous pouvez faire appel au souvenir de vos expériences réelles pour évaluer dans quelle mesure la reproduction correspond au son réel original. Pour d'autres sortes de musique, par exemple musique légère ou électronique, vous devrez imaginer ce que l'on voulait vous faire entendre et évaluer dans quelle mesure la reproduction correspond au son voulu.»

Note. — Pour éviter toute confusion sur celui des critères ci-dessus qu'il fallait utiliser, l'expérimentateur peut suggérer aux auditeurs celui des programmes reproduits qu'il convient d'évaluer en termes de fidélité au son réel original, et celui qu'il convient d'évaluer en termes de fidélité au son «voulu».

«Dans tous les cas, les évaluations de la fidélité sont faites au moyen d'une échelle graduée de 0 à 10 comme indiqué ci-dessous :



«Tout nombre compris entre 0 et 10, comportant une décimale, peut être utilisé pour déterminer le degré de fidélité des reproductions. Chacun des deux haut-parleurs dans chaque présentation par paires devra être évalué sur la même base. Ecoutez attentivement les deux présentations avant de noter vos deux évaluations. Quelle que soit la similitude des sons reproduits par certains haut-parleurs appairés, vous ne devez pas donner la même note à leurs évaluations; des nombres différents doivent être utilisés. Enfin, tâchez de ne jamais oublier que vos préférences musicales ne devraient pas entrer en ligne de compte dans les évaluations qui se rapportent uniquement à la fidélité de la reproduction sonore».

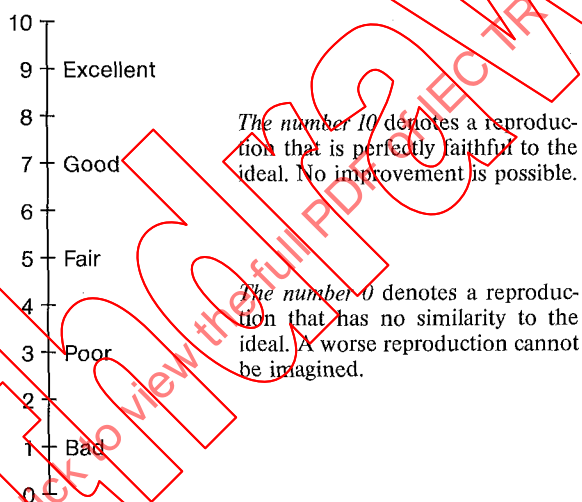
Note. — Il convient que les instructions donnent également des informations sur la durée approximative de l'essai, les pauses, les essais préliminaires et d'autres détails utiles aux auditeurs.

A2. Pairwise comparisons

“In this experiment you will hear short programmes of music and speech reproduced over a number of loudspeakers. Each programme will be presented twice in rapid succession, each time over a different loudspeaker. The order of the programmes and the combination of loudspeakers in the pairs vary in a random way. At the end of each paired presentation you should judge the reproduction from each of the two loudspeakers in terms of its fidelity to an ideal. For some music and speech, you can use your memory of live experiences to rate how well the reproduction matches the original live sound. For other music, for example certain popular music and electronic music, you will have to imagine what was intended for you to hear and rate how well the reproduction matches the intended sound.”

Note. — To avoid possible confusion about which of the above criteria should be used, the experimenter may suggest to the listener(s) which of the reproduced programmes should be rated in terms of fidelity to the original live sound, and which should be rated in terms of fidelity to the “intended” sound.

“In all cases the fidelity ratings shall be made on a scale from 0 to 10, as shown below:



“Any number between 0 and 10, including one decimal place, can be used to characterize the degree of fidelity in the reproductions. Each of the two loudspeakers in each paired presentation should be rated on the same basis. Listen carefully to both presentations before writing down your two ratings. However similar some paired loudspeakers may sound you must not give them the same number rating; different numbers must be used. Finally, try to remember that your musical preferences should not enter into the ratings, which refer only to the fidelity of the sound reproduction.”

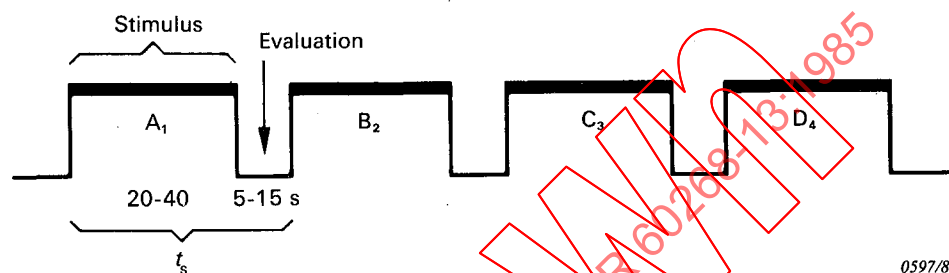
Note. — The instruction should also give information about the approximate test duration, pauses, preliminary trials and other details helpful to the listeners.

ANNEXE B

COMPARAISON DE LA LONGUEUR D'ESSAI

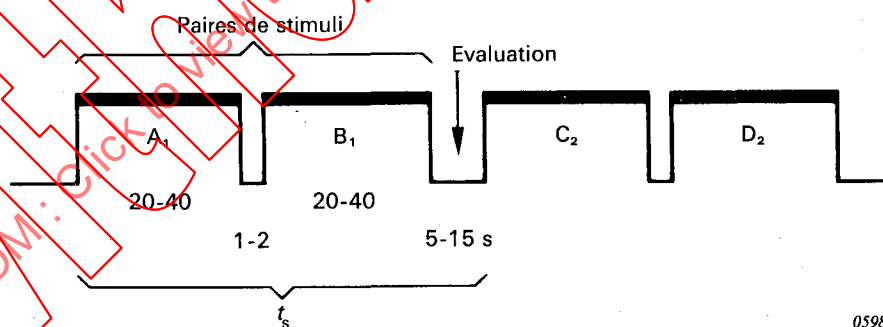
Les diagrammes des temps de présentation des stimuli, qui résultent des deux méthodes expérimentales définies au paragraphe 3.4, de la durée des programmes et des pauses recommandés au paragraphe 3.2, peuvent être illustrés par les exemples suivants:

I) Evaluation par stimulus isolé (S)



Un stimulus est un programme donné (1, 2, 3, 4, etc.) reproduit par un haut-parleur donné (A, B, C, D, etc.); Exemple: A₁ signifie haut-parleur A reproduisant le programme 1.

II) Comparaison par paires (P)



Une paire de stimuli est un programme donné (1, 2, 3, 4, etc.) reproduit par deux haut-parleurs différents (A et B, C et D, etc.).

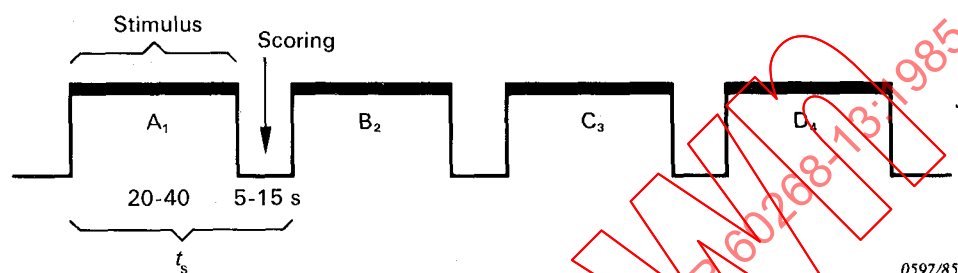
Dans la présentation suivante, une comparaison est faite entre les durées d'essai pour les méthodes S et P. A partir de ces résultats, les utilisateurs peuvent choisir la méthode la plus efficace pour des usages spécifiques.

APPENDIX B

COMPARISON OF TEST LENGTH

The time patterns for stimulus presentation, according to the two experimental procedures defined in Sub-clause 3.4 and to the duration of programme sections and pauses recommended in Sub-clause 3.2, can be illustrated by the following examples:

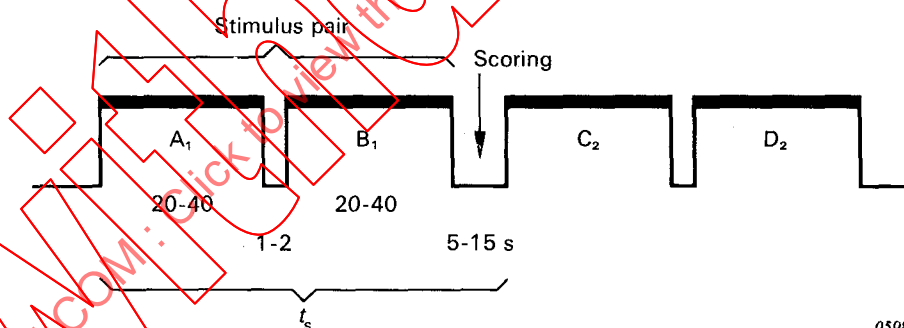
I) Single stimulus rating (S)



0597/85

A stimulus comprises a given programme (1, 2, 3, 4, etc.) reproduced by a given loudspeaker (A, B, C, D, etc.); Example: A₁ denotes loudspeaker A reproducing programme 1.

II) Paired comparison (P)



0598/85

A stimulus pair comprises a given programme (1, 2, 3, 4, etc.) reproduced by two different loudspeakers (A and B, C and D, etc.).

In the following presentation a comparison is made between the test durations for the S and P procedures. From this data, users can choose the method that is more efficient for specific purposes.

Notations:

n = nombre de haut-parleurs

m = nombre de programmes

r = nombre de répétitions

t_i = durée des instructions

t_p = durée des essais préliminaires

Suppositions: $t_i = 7$ min

$t_p = (n + m)$ minutes

N = nombre de présentations (note 1)

t_s = durée de présentation d'un programme, y compris les pauses

$T = N \cdot t_s$ = durée spécifique d'écoute par auditeur et par programme (tableau BII)

$L = T \cdot m + t_i + t_p$ = durée totale de l'essai par auditeur (tableau BIII).

(Toutes les durées ne sont que des exemples, choisis dans les limites de durées permises, et servent surtout de moyen de comparaison.)

TABLEAU BI

Calcul de la durée de présentation t_s (en secondes)

Durée	S	P
Stimulus A ₁	30	30
Pause	—	2
Stimulus B ₁	—	30
Pause	15	15
t_s	45	77 (par paire)

TABLEAU BII

Calcul de la durée spécifique d'écoute T (minutes par auditeur et par programme)

n	$n-1$	r^{**}	N présentations*		T min ($N \times t_s$)	
			S ($n \times r$)	P $\frac{1}{2} n(n-1)$	S	P
1	0	4	4	—	3	—
2	1	3	6	3	4,5	3,85
3	2	4	12	6	9	7,7
4	3	3	12	6	9	7,7
5	4	4	20	10	15	12,85
6	5	4	24	15	18	19,25
...	...	...	...	...	...	...
10	9	4	40	45	30	57,75

* Voir note 1, page 42.

** Voir note 2, page 42.

Given values:

n = number of loudspeakers

m = number of programmes

r = number of replications

t_i = duration of instruction

t_p = duration of preliminary trials

Assumptions: $t_i = 7$ min

$t_p = (n + m)$ minutes

N = number of presentations (Note 1)

t_s = duration of presentation of a programme section, including pauses

$T = N \cdot t_s$ = specific listening time per listener and programme (Table BII)

$L = T \cdot m + t_i + t_p$ = total test length per listener (Table BIII).

(All time values are examples only, chosen within the permitted range of durations, and serve mainly for the purpose of comparison.)

TABLE BI

Calculation of the presentation time t_s (in seconds)

Duration	S	P
Stimulus A ₁	30	30
Pause	—	2
Stimulus B ₁	—	30
Pause	15	15
t_s	45	77 (per pair)

TABLE BII

Calculation of the specific listening time T (minutes per listener and per programme)

n	$n-1$	r^{**}	N presentations*		T min ($N \times t_s$)	
			S ($n \times r$)	P $\frac{1}{2} n(n-1)$	S	P
1	0	4	4	—	3	—
2	1	3	6	3	4.5	3.85
3	2	4	12	6	9	7.7
4	3	3	12	6	9	7.7
5	4	4	20	10	15	12.85
6	5	4	24	15	18	19.25
...	...	...	...	...	...	...
10	9	4	40	45	30	57.75

* See Note 1, page 43.

** See Note 2, page 43.

TABLEAU BIII

Calcul de la durée totale d'essai L(en minutes, arrondies au plus près) pour $m = 5$ programmes

n	t_p ($n + m$)	$L \text{ min}$ ($mT + t_i + t_p$)	
		S	P
1	6	28	—
2	7	37	34
3	8	60	54
4	9	61	55
5	10	92	82
6	11	108	115
·	·	·	·
·	·	·	·
10	15	172	311

Notes 1. — Le nombre de présentations N est égal à $n \cdot r$ stimuli pour des stimuli isolés (S) ou à $n(n-1)/2$ paires de stimuli dans la comparaison par paires (P), à l'exception de la répétition nécessaire lorsque $n < 4$ dans la méthode par paires (P).

2. — La comparaison dépend évidemment de la valeur de r . Pour $2 \leq n \leq 5$, r a été choisi de telle sorte que dans la méthode P, le nombre de répétitions minimal ($n-1$) puisse être multiplié par un nombre entier (3, 2 ou 1) pour obtenir le nombre requis de répétitions (3 ou 4). Pour $n > 5$, ($n-1$) sera toujours supérieur à r et la comparaison entre S et P ne se fait plus sur un pied d'égalité, du fait que le nombre plus grand de répétitions pour P augmentera en même temps la fiabilité. L'essai plus long sera donc également plus précis.

Il ressort du tableau BIII qu'avec les durées, les répétitions et le nombre de programmes pris comme exemples, la méthode P aboutit à une longueur totale d'essai plus courte lorsque cinq haut-parleurs au maximum sont utilisés dans l'essai. Pour six haut-parleurs ou plus dans le même essai, l'essai S est plus court.

TABLE BIII

Calculation of the total test length L
(in minutes, rounded off) for $m = 5$ programmes

n	t_p ($n+m$)	L min ($mT + t_i + t_p$)	
		S	P
1	6	28	—
2	7	37	34
3	8	60	54
4	9	61	55
5	10	92	82
6	11	108	115
...	...	...	...
10	15	172	311

Notes 1. — The number of presentations N is obtained as $n \cdot r$ single stimuli, or $n(n-1)/2$ pairs, for S and P respectively, except for the necessary multiplication for P when $n \leq 4$.

2. — The comparison is of course dependent on the value of r . For $2 \leq n \leq 5$, r has been chosen in such a way that for P the minimum number of replications ($n-1$) can be multiplied by an integer (3, 2 or 1) to make up the desired number of replications (3 or 4). For $n > 5$, ($n-1$) will always be greater than r , and the comparison between S and P is no longer on equal terms, since the larger number of replications for P at the same time will increase the reliability. The longer test will therefore also be more accurate.

It can be concluded from Table BIII that with the assumed durations, replications and number of programmes, the P procedure leads to a shorter total test length when up to five loudspeakers are included in the test. For six or more loudspeakers in the same test, the S test is shorter.

ANNEXE C

TRAITEMENT STATISTIQUE DES RÉSULTATS À PARTIR DES ESSAIS D'ÉCOUTE

Sommaire:

- C1. Introduction
- C2. Tableaux de notes, statistique descriptive
 - C2.1 Tableau de notes individuelles (exemple)
 - C2.2 Tableau de notes d'un groupe
- C3. Analyse de variance (AN.VAR)
 - C3.1 AN.VAR pour tableau de notes individuelles
 - C3.2 AN.VAR pour tableau de résultats de groupe
- C4. Essais pour les comparaisons spécifiques
 - C4.1 Comparaisons indépendantes planifiées
 - C4.2 Comparaisons dépendantes planifiées
 - C4.3 Comparaisons non planifiées
 - C4.4 Comparaisons spécifiques pour des programmes isolés
- C5. Essais d'écoute comportant des variables supplémentaires
- C6. Fiabilité
 - C6.1 Fiabilité individuelle
 - C6.2 Fiabilité interindividuelle
- C7. Quelques points critiques dans la validité des essais
- C8. Calculs, programmes d'ordinateur, tableaux
- C9. Présentation des résultats
- C10. Conclusions
- C11. Références

Figure C1

C1. Introduction

La présente annexe est une version abrégée d'un rapport technique détaillé sur le traitement statistique des résultats à partir des essais d'écoute (Gabrielsson, 1979). L'annexe se borne à présenter les principes généraux de l'analyse statistique définie dans ce rapport. En ce qui concerne les détails d'application et la compréhension plus complète de l'analyse, il est recommandé de consulter le rapport technique cité ci-dessus et/ou la bibliographie qu'il donne.

Au cours d'un essai d'écoute, chaque auditeur est supposé faire un certain nombre d'évaluations sur l'échelle de «fidélité», graduée de 0 à 10 pour chaque combinaison, au cours de l'essai, d'un haut-parleur avec un programme (comme indiqué dans l'article 3).

Pour que les notes donnent autant d'information que possible, il peut être préférable de les traiter individuellement (c'est-à-dire pour chaque auditeur) aussi bien qu'interindividuellement (c'est-à-dire pour l'ensemble des auditeurs dans le cadre du même groupe). Le traitement statistique peut être divisé sans inconvénient en deux niveaux:

APPENDIX C

STATISTICAL TREATMENT OF DATA FROM LISTENING TESTS

Contents:

- C1. Introduction
- C2. Data matrices, descriptive statistics
 - C2.1 Individual data matrix (example)
 - C2.2 Group data matrix
- C3. Analysis of variance (ANOVA)
 - C3.1 ANOVA for individual data matrix
 - C3.2 ANOVA for group data matrix
- C4. Tests for specific comparisons
 - C4.1 Planned independent comparisons
 - C4.2 Planned non-independent comparisons
 - C4.3 Non-planned comparisons
 - C4.4 Specific comparisons within single programmes
- C5. Listening tests including extra variables
- C6. Reliability
 - C6.1 Intra-individual reliability
 - C6.2 Inter-individual reliability
- C7. Some critical issues in significance testing
- C8. Computations, computer programmes, tables
- C9. Presentation of results
- C10. Concluding comments
- C11. References

Figure C1

C1. Introduction

This appendix is an abbreviated version of a detailed technical report on the statistical treatment of data from listening tests (Gabrielsson, 1979). The appendix is limited to the presentation of the general principles for the statistical analysis described in this report. For details of applications and for more complete understanding of the analysis it is recommended to consult the above-mentioned technical report and/or literature cited therein.

Each listener in a listening test is assumed to make a certain number of ratings on the 0 to 10 "fidelity" scale for each loudspeaker programme combination in the test (as described in Clause 3).

To get as much information as possible from the data it may be preferable to treat them intra-individually (that is, within each listener) as well as inter-individually (that is, over listeners within the same group of listeners). The statistical treatment may conveniently be divided into two steps:

- 1) *Statistiques descriptives.* A ce stade, les notes d'évaluation sont inscrites dans les tableaux appropriés pour faciliter leur suivi et on calcule certaines fonctions statistiques courantes (par exemple, les moyennes arithmétiques). Les notes peuvent également se présenter sous forme graphique. L'examen visuel des tableaux, des graphes et des statistiques calculées conduit généralement à certaines conclusions concernant les haut-parleurs en essai.

Il convient d'utiliser toujours ce type de statistique descriptive qui, en général, ne présente pas de difficulté particulière. Il est décrit dans l'article C2 et dans certaines parties de l'article C9.

- 2) *Statistiques par induction.* A ce stade, l'analyse des notes d'évaluation est approfondie au moyen de l'analyse de variance (AN.VAR) et des méthodes associées pour vérifier si les différences d'évaluation entre différents haut-parleurs (et/ou entre différents programmes) ont ou n'ont pas de signification statistique, et s'il existe une interaction entre haut-parleurs et programmes. L'AN.VAR peut être également utilisée pour estimer la fiabilité des notes individuelles et interindividuelles.

L'application de ce type de traitement statistique est facultative. Elle demande plus de travail et plus de connaissances des statistiques. Par ailleurs, elle permet généralement à l'utilisateur d'obtenir des informations plus détaillées à partir des notes et d'aboutir à des conclusions plus précises. Elle s'applique aussi facilement à des essais d'écoute plus complexe. Elle est décrite dans les articles C3 à C8.

Afin d'illustrer ces deux niveaux, on donnera un exemple utilisant des notes réelles à partir d'un essai d'écoute utilisant quatre haut-parleurs et cinq programmes. L'exemple se rapporte aux évaluations sur une échelle de «fidélité» graduée de 0 à 10. Toutefois, les procédures statistiques décrites ci-dessous peuvent également être utilisées pour d'autres échelles d'évaluation («clarté», «plénitude», «intensité sonore» ou autres cas similaires) construites de façon analogue.

C2. Tableaux de notes, statistique descriptive

C2.1 Tableau de notes individuelles (exemple)

Pour un auditeur donné dans cet essai d'écoute, on a obtenu les résultats suivants, voir tableau CI. Il a fait trois évaluations pour chaque combinaison de haut-parleur avec un programme. Ces trois notes figurent dans le rang supérieur de chaque case et leur moyenne arithmétique (M) est indiquée juste au-dessous.

TABLEAU CI
Exemple d'un tableau de notes individuelles

		Haut-parleur				Moyenne par programme
		A	B	C	D	
Programme	1	7 6 7 M = 6,7	5 5 5 5,0	6 7 5 6,0	3 3 4 3,3	5,3
	2	6 6 7 6,3	3 3 4 3,3	5 5 7 5,7	3 3 4 3,3	4,7
	3	7 8 8 7,7	2 2 2 2,0	7 7 7 7,0	3 3 3 3,0	4,9
	4	7 8 8 7,7	3 3 3 3,0	8 7 8 7,7	3 3 3 3,0	5,3
	5	6 7 6 6,3	5 5 4 4,7	6 6 6 6,0	5 5 5 5,0	5,5
Moyenne par haut-parleur		6,9	3,6	6,5	3,5	

- 1) *Descriptive statistics.* In this step the rating data are entered into suitable matrices to make them easily surveyable, and certain common statistics (e.g. arithmetic means) are computed. The data may also be displayed in graphical form. Visual inspection of the matrices, the graphs, and the computed statistics usually lead to certain conclusions about the loudspeakers under test.

This type of descriptive statistics should always be applied and generally presents no special difficulties. It is described in Clause C2 and in parts of Clause C9.

- 2) *Interference statistics.* In this step the rating data are analyzed further by means of analysis of variance (ANOVA) and related procedures to test if differences between the ratings for different loudspeakers (and/or for different programmes) are statistically significant or not, and if there is an interaction between loudspeakers and programmes. ANOVA may also be used to estimate the reliability of the data, intra-individually and inter-individually.

The application of this type of statistical treatment is optional. It requires more work and more knowledge about statistics. On the other hand it usually enables the user to extract more detailed information from the data and to arrive at more definite conclusions. It is also easily generalized to more complex listening tests. It is described in Clauses C3 to C8.

In order to illustrate these steps an example will be given using real data from a listening test with four loudspeakers and five programmes. The example refers to ratings on a 0 to 10 "fidelity" scale. However, the statistical procedures described below may also be used for other rating scales ("brightness", "fullness", "loudness", etc.) constructed in similar ways.

C2. Data matrices, descriptive statistics

C2.1 Individual data matrix (example)

For a certain listener in this listening test the following data were obtained, see Table CI. He made three ratings per each loudspeaker $\times$ programme combination. These three values appear in the upper row of each cell, and their arithmetic mean (M) is given directly below.

TABLE CI
Example of individual data matrix

		Loudspeaker				Means for programmes
		A	B	C	D	
Programme	1	7 6 7 M = 6.7	5 5 5 5.0	6 7 5 6.0	3 3 4 3.3	5.3
	2	6 6 7 6.3	3 3 4 3.3	5 5 7 5.7	3 3 4 3.3	4.7
	3	7 8 8 7.7	2 2 2 2.0	7 7 7 7.0	3 3 3 3.0	4.9
	4	7 8 8 7.7	3 3 3 3.0	8 7 8 7.7	3 3 3 3.0	5.3
	5	6 7 6 6.3	5 5 4 4.7	6 6 6 6.0	5 5 5 5.0	5.5
Means for loudspeakers		6.9	3.6	6.5	3.5	

L'examen visuel direct de ce tableau révèle de nombreux points concernant les résultats. Cet auditeur fait preuve d'une grande stabilité dans ses cotations (grande fiabilité individuelle) — les trois notes à l'intérieur de chaque case sont à peine différentes ou même identiques dans de nombreux cas. Les notes moyennes des haut-parleurs (sur cinq programmes) figurent dans la ligne en bas du tableau et indiquent que les haut-parleurs A et C sont supérieurs aux haut-parleurs B et D. Les notes moyennes pour chaque programme (reproduits par les quatre haut-parleurs) figurent dans la marge de droite de ce tableau et suggèrent que les programmes 2 et 3 sont plus difficiles à reproduire que les autres. En examinant les moyennes par haut-parleur pour chaque programme, il est facile de constater que les différences entre les haut-parleurs varient d'un programme à l'autre et, parfois, s'éloignent beaucoup des différences correspondantes relevées sur les moyennes de bas de page. Par exemple, la différence moyenne entre les haut-parleurs A et B, pour tous les cinq programmes, est 3,3 (6,9 – 3,6), alors que cette différence tombe à 1,6 (6,3 – 4,7) pour le programme 5 et monte à 5,7 (7,7 – 2,0) pour le programme 3. Cela suggère qu'il faut approfondir l'interaction entre les haut-parleurs et les programmes.

C2.2 *Tableau de notes d'un groupe*

Le tableau de notes d'un groupe d'auditeurs est constitué par une combinaison des tableaux individuels. Il peut être représenté de bien des façons différentes. Une de ces façons est explicitée par le tableau CII pour un groupe de quatre auditeurs désignés S, T, U et V. (Les notes correspondant à l'auditeur S sont les mêmes que celles du tableau CI, ces quatre auditeurs ont été sélectionnés dans un groupe plus grand comportant dix personnes.)

Pour chaque combinaison d'un haut-parleur avec un programme, il y a douze notes, trois par auditeur. M_S est la moyenne arithmétique des trois évaluations de l'auditeur S, M_T indique celle de T et ainsi de suite. M_g (g pour groupe) indique la moyenne arithmétique pour tout le groupe d'auditeurs. (Ces désignations ne figurent que dans la case supérieure gauche, mais sont implicites dans les autres cases.)

Les moyennes par haut-parleur dans la ligne en bas de page représentent les notes moyennes des haut-parleurs établies sur l'ensemble des programmes et des auditeurs. Les moyennes par programme dans la marge de droite représentent les notes moyennes pour les différents programmes, établies sur l'ensemble des haut-parleurs et des auditeurs.

L'examen visuel de ce tableau aboutit à peu près aux mêmes conclusions que l'examen du tableau CI (qui représentait le tableau des notes individuelles pour l'auditeur S): les haut-parleurs A et C sont supérieurs aux haut-parleurs B et D, une interaction apparaît entre les haut-parleurs et les programmes, et ainsi de suite. Ce tableau permet également d'étudier les différences individuelles de notes et/ou des interactions entre auditeurs et haut-parleurs.

Direct visual inspection of this matrix reveals much about the results. This listener shows a high stability in his ratings (high intra-individual reliability) and the three ratings within each cell differ very little or are even the same in many cases. The mean ratings for the loudspeakers (over the five programmes) appear in the bottom margin and indicate that loudspeakers A and C are superior to loudspeakers B and D. The mean ratings for each programme (over the four loudspeakers) appear in the right-hand margin and suggest that programmes 2 and 3 are harder to reproduce than the other programmes. Looking at the means for the loudspeakers within each programme it is easily seen that the differences between the loudspeakers vary from programme to programme and sometimes differ significantly from the corresponding differences between the means in the bottom margin. For instance, the average difference between loudspeakers A and B over all five programmes is 3.3 (6.9 - 3.6), while the difference is as small as 1.6 (6.3 - 4.7) for programme 5 and as big as 5.7 (7.7 - 2.0) for programme 3. This suggests an interaction between loudspeakers and programmes to be further studied.

C2.2 Group data matrix

The data matrix for a group of listeners is constituted by a combination of individual data matrices. It may be represented in many different ways. One way is illustrated in Table CII for a group of four listeners, S, T, U and V. (The data for listener S are the same as in Table CI; these four listeners were taken from a larger group of ten listeners.)

For each loudspeaker $\times$ programme combination there are twelve ratings, three per listener. M_S denotes the arithmetic mean of the three ratings by listener S, M_T the mean for listener T, and so on. M_g (g for group) denotes the arithmetic mean for the whole group of listeners. (These designations are written only in the upper left-hand box but are implicit in the other boxes.)

The means for loudspeakers in the bottom margin represent the mean ratings for the loudspeakers, averaged over programmes and listeners. The means for programmes in the right-hand margin represent the mean ratings at the different programmes, averaged over loudspeakers and listeners.

Visual inspection of this matrix leads to much the same conclusions as the inspection of Table CI (which was the individual data matrix for listener S): loudspeakers A and C are superior to loudspeakers B and D, there are suggestions of interaction between loudspeakers and programmes, and so on. This matrix also makes it possible to study individual differences in the ratings and/or interactions between listeners and loudspeakers.

TABLEAU CII

Exemple de tableau de résultats de groupe pour les auditeurs désignés S, T, U et V

Pro-gramme	Audi-teur	Haut-parleur								Moyenne par pro-gramme
		A		B		C		D		
1	S	7 6 7	$M_S = 6,7$	5 5 5	5,0	6 7 5	6,0	3 3 4	3,3	5,3
	T	5 7 4	$M_T = 5,3$	4 4 3	3,7	5 5 5	5,0	3 3 3	3,0	
	U	6 7 7	$M_U = 6,7$	4 5 4	4,3	5 7 7	6,3	5 5 5	5,0	
	V	8 7 7	$M_V = 7,3$	7 6 5	6,0	6 6 7	6,3	7 4 5	5,3	
	$M_g = 6,5$		4,8		5,9		4,2			
2	S	6 6 7	6,3	3 3 4	3,3	5 5 7	5,7	3 3 4	3,3	5,0
	T	8 7 8	7,7	3 2 3	2,7	4 4 7	5,0	4 3 2	3,0	
	U	6 8 8	7,3	4 3 3	3,3	6 7 8	7,0	3 3 3	3,0	
	V	7 8 7	7,3	4 4 5	4,3	8 7 4	6,3	4 4 4	4,0	
	7,2		3,4		6,0		3,3			
3	S	7 8 8	7,7	2 2 2	2,0	7 7 7	7,0	3 3 3	3,0	4,7
	T	5 4 5	4,7	3 2 2	2,3	5 3 3	3,7	1 1 2	1,3	
	U	7 7 9	7,7	4 4 3	3,7	7 7 9	7,7	3 4 4	3,7	
	V	9 6 6	7,0	5 4 6	5,0	8 4 6	6,0	3 3 3	3,0	
	6,8		3,3		6,1		2,8			
4	S	7 8 8	7,7	3 3 3	3,0	8 7 8	7,7	3 3 3	3,0	5,3
	T	6 7 8	7,0	4 3 3	3,3	5 4 5	4,7	3 1 2	2,0	
	U	6 7 8	7,0	3 4 3	3,3	6 6 7	6,3	1 4 2	2,3	
	V	8 8 7	7,7	4 6 5	5,0	8 9 8	8,3	7 7 4	6,0	
	7,3		3,7		6,8		3,3			
5	S	6 7 6	6,3	5 5 4	4,7	6 6 6	6,0	5 5 5	5,0	5,1
	T	6 5 7	6,0	4 3 4	3,7	5 6 3	4,7	4 3 2	3,0	
	U	6 5 6	5,7	3 3 3	3,0	4 5 5	4,7	5 4 4	4,3	
	V	9 7 7	7,7	4 4 5	4,3	7 7 5	6,3	6 6 5	5,7	
	6,4		3,9		5,4		4,5			
Moyenne par haut-parleur		6,8		3,8		6,0		3,6		

TABLE CII

Example of group data matrix for listeners denoted S, T, U and V

Pro- gramme	Lis- tener	Loudspeaker								Means for pro- grammes
		A		B		C		D		
1	S	7 6 7	$M_S = 6.7$	5 5 5	5.0	6 7 5	6.0	3 3 4	3.3	5.3
	T	5 7 4	$M_T = 5.3$	4 4 3	3.7	5 5 5	5.0	3 3 3	3.0	
	U	6 7 7	$M_U = 6.7$	4 5 4	4.3	5 7 7	6.3	5 5 5	5.0	
	V	8 7 7	$M_V = 7.3$	7 6 5	6.0	6 6 7	6.3	7 4 5	5.3	
	$M_g = 6.5$		4.8		5.9		4.2			
2	S	6 6 7	6.3	3 3 4	3.3	5 5 7	5.7	3 3 4	3.3	5.0
	T	8 7 8	7.7	3 2 3	2.7	4 4 7	5.0	4 3 2	3.0	
	U	6 8 8	7.3	4 3 3	3.3	6 7 8	7.0	3 3 3	3.0	
	V	7 8 7	7.3	4 4 5	4.3	8 7 4	6.3	4 4 4	4.0	
	7.2		3.4		6.0		3.3			
3	S	7 8 8	7.7	2 2 2	2.0	7 7 7	7.0	3 3 3	3.0	4.7
	T	5 4 5	4.7	3 2 2	2.3	5 3 3	3.7	1 1 2	1.3	
	U	7 7 9	7.7	4 4 3	3.7	7 7 9	7.7	3 4 4	3.7	
	V	9 6 6	7.0	5 4 6	5.0	8 4 6	6.0	3 3 3	3.0	
	6.8		3.3		6.1		2.8			
4	S	7 8 8	7.7	3 3 3	3.0	8 7 8	7.7	3 3 3	3.0	5.3
	T	6 7 8	7.0	4 3 3	3.3	5 4 5	4.7	3 1 2	2.0	
	U	6 7 8	7.0	3 4 3	3.3	6 6 7	6.3	1 4 2	2.3	
	V	8 8 7	7.7	4 6 5	5.0	8 9 8	8.3	7 7 4	6.0	
	7.3		3.7		6.8		3.3			
5	S	6 7 6	6.3	5 5 4	4.7	6 6 6	6.0	5 5 5	5.0	5.1
	T	6 5 7	6.0	4 3 4	3.7	5 6 3	4.7	4 3 2	3.0	
	U	6 5 6	5.7	3 3 3	3.0	4 5 5	4.7	5 4 4	4.3	
	V	9 7 7	7.7	4 4 5	4.3	7 7 5	6.3	6 6 5	5.7	
	6.4		3.9		5.4		4.5			
Means for loudspeakers		6.8		3.8		6.0		3.6		

Le tableau CIII présente des résultats de groupe d'une manière plus condensée. Les notes individuelles sont omises et seules les moyennes arithmétiques pour chaque combinaison (programme $\times$ haut-parleur) sont indiquées (les valeurs M_g du tableau CII) en même temps que les moyennes par haut-parleur ou par programme dans les marges.

TABLEAU CIII

Tableau condensé de résultats de groupe

		Haut-parleur				Moyenne par programme
		A	B	C	D	
Programme	1	6,5	4,8	5,9	4,2	5,3
	2	7,2	3,4	6,0	3,3	5,0
	3	6,8	3,3	6,1	2,8	4,7
	4	7,3	3,7	6,8	3,3	5,3
	5	6,4	3,9	5,4	4,5	5,1
Moyenne par haut-parleur		6,8	3,8	6,0	3,6	

C3. Analyse de variance (AN.VAR)

Un examen visuel des tableaux de résultats peut parfois donner des informations suffisantes pour les objectifs recherchés. Toutefois, pour obtenir des informations plus détaillées à partir des résultats et pour aboutir à des conclusions plus précises, il est préférable d'utiliser des méthodes statistiques telles que l'analyse de variance (AN.VAR) et les procédures associées pour tester la fiabilité. L'AN.VAR prouve essentiellement que la variance totale dans les résultats est formée des différentes composantes dues aux différentes causes de variation dans un essai d'écoute, c'est-à-dire haut-parleurs, programmes, auditeurs, et interactions éventuelles entre ces variables. Les essais statistiques associés permettent de déterminer, avec une probabilité donnée, si les différences de notes entre haut-parleurs et/ou programmes sont des différences « vraies » ou si elles risquent d'être dues au hasard. De même, il est possible de déterminer s'il existe de « véritables » interactions ou pas.

Les raisonnements et les hypothèses servant de base à ces procédures sont discutés dans la plupart des textes sur les statistiques et la méthode expérimentale (par exemple, Hays, 1973; Kirk, 1968, 1972; Winer, 1971).

C3.1 AN.VAR pour tableau de notes individuelles

Une AN.VAR des résultats du tableau CI peut être réalisée facilement par n'importe lequel des nombreux programmes d'ordinateur disponibles pour AN.VAR (voir article C8). Les résultats sont présentés dans un tableau récapitulatif de la façon suivante:

A more condensed way of presenting the group data is shown in Table CIII. All individual values are omitted, and only the arithmetic means for each programme $\times$ loudspeaker combination are given (the M_g values in Table CII) together with the means for the loudspeakers and for programmes in the margins.

TABLE CIII
Condensed group data matrix

		Loudspeaker				Means for programmes
		A	B	C	D	
Programme	1	6.5	4.8	5.9	4.2	5.3
	2	7.2	3.4	6.0	3.3	5.0
	3	6.8	3.3	6.1	2.8	4.7
	4	7.3	3.7	6.8	3.3	5.3
	5	6.4	3.9	5.4	4.5	5.1
Means for loudspeakers		6.8	3.8	6.0	3.6	

C3. Analysis of variance (ANOVA)

Visual inspection of data matrices may sometimes give sufficient information for the purposes at hand. However, to be able to extract more detailed information from the data and arrive at more definite conclusions it may be preferable to use statistical methods like analysis of variance (ANOVA) and related procedures for significance testing. ANOVA essentially means that the total variance in the data is split up into different components due to the different sources of variation in a listening test, that is, loudspeakers, programmes, listeners, and possible interactions between these variables. The related statistical tests make it possible to decide, with a certain probability, whether the differences in ratings between different loudspeakers and/or different programmes are "true" differences, or if they may be due to chance. Similarly it is possible to decide whether there are some "true" interactions or not.

The rationale and the assumptions underlying these procedures are discussed in most texts on statistics and experimental design (for instance, Hays, 1973; Kirk, 1968, 1972; Winer, 1971).

C3.1 ANOVA for individual data matrix

An ANOVA on the data in Table CI may be conveniently performed by any of many available computer programmes for ANOVA (see Clause C8). The results are presented in a summary table as follows:

TABLEAU CIV

Exemple du tableau récapitulatif pour AN.VAR effectuée sur un tableau de résultats individuels

Source de variations	Somme de carrés (SS)	Degrés de liberté (df)	Moyenne des carrés (MS)	F	p
Haut-parleurs (L)	148,93	3	49,64	177,29	< 0,01
Programmes (P)	5,43	4	1,36	4,86	< 0,01
L × P	35,23	12	2,94	10,50	< 0,01
à l'intérieur de la case	11,33	40	0,28		
Total	200,92	59			

La moyenne des carrés (MS) pour la variation «à l'intérieur d'une case» reflète la variance des notes à l'intérieur de toutes les cases du tableau CI et est considérée comme une estimation de la «variance d'erreur» pour cet auditeur. Comme l'indique le tableau CI, cet auditeur était très stable dans ses évaluations et la variance d'erreur estimée se réduit ainsi à 0,28. La moyenne des carrés (MS) pour les haut-parleurs (49,64) représente une estimation de la variance due aux différences entre les haut-parleurs (les différences entre les moyennes des haut-parleurs dans la ligne en bas du tableau CI) plus une estimation de la variance d'erreur. En conséquence, plus grande est la moyenne des carrés des haut-parleurs par rapport à la moyenne des carrés à l'intérieur de la case (qui ne représente qu'une variance d'erreur), plus il est probable qu'il existe des différences «vraies» (non aléatoires) entre les haut-parleurs (comme évalué par cet auditeur). Cela est formellement vérifié par l'essai F qui signifie, dans ce cas, que la moyenne des carrés du haut-parleur, divisée par la moyenne des carrés à l'intérieur de la case, donne un rapport F de 177,29 ($= 49,64/0,28$). Cette valeur est à comparer avec la valeur F dite «critique» pour les degrés de liberté respectifs (à savoir $df = 3$ pour les haut-parleurs et 40 pour la variation à l'intérieur de la case) comme indiqué dans les tableaux de la loi de F , qui se trouvent dans la plupart des manuels sur les statistiques. En utilisant le niveau de signification 0,01 (1%), la valeur critique F pour ces valeurs de df est 4,31. La valeur F obtenue (177,29) est bien supérieure à la valeur critique, ce qui la rend significative au niveau 0,01 (1%). Dans la colonne droite du tableau CIV, cela est désigné par le symbole $p < 0,01$, ce qui veut dire que la probabilité (p) d'obtenir les différences observées entre les haut-parleurs, uniquement par hasard, est inférieure à 0,01. En conséquence, les différences peuvent être considérées comme «vraies».

Un raisonnement analogue peut être fait en ce qui concerne les programmes et l'interaction (haut-parleurs × programmes). Comme l'indique le tableau CIV, les essais F correspondants sont également significatifs au niveau 0,01.

Tel qu'il est utilisé dans ce cas, l'essai F fonctionne comme un essai «global». Par exemple, un rapport F significatif en ce qui concerne les haut-parleurs indique seulement qu'il existe au moins une différence significative parmi toutes les combinaisons possibles des haut-parleurs (c'est-à-dire, la différence significative peut être celle entre A et B, et/ou entre A et C, et/ou entre A et D, B et C, B et D, C et D — et/ou entre des combinaisons plus complexes, par exemple la moyenne de A et C contre la moyenne B et D). Pour trouver exactement quelle différence est significative, il est nécessaire de réaliser des essais pour comparaisons spécifiques, voir article C4. D'autre part, un rapport F non significatif pour les haut-parleurs voudrait dire qu'il n'existe aucune différence significative entre les haut-parleurs dans n'importe quelle combinaison.

L'interprétation d'une interaction significative (haut-parleurs × programmes) peut être faite souvent d'une façon valable par examen direct du tableau de résultats. Les essais statistiques à cet effet sont mentionnés dans l'article C4. Il est important d'étudier le sens d'une interaction significative, parce qu'elle peut donner des informations intéressantes sur la qualité des différents haut-parleurs pour différents types du programme convenu.

TABLE CIV

Example of summary table for ANOVA on individual data matrix

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F	p
Loudspeakers (L)	148.93	3	49.64	177.29	<0.01
Programmes (P)	5.43	4	1.36	4.86	<0.01
L × P	35.23	12	2.94	10.50	<0.01
Within-cell	11.33	40	0.28		
Total	200.92	59			

The mean square (MS) for the “within-cell” variation reflects the variance of the ratings within all cells of Table CI and is taken as an estimate of the “error variance” for this listener. As noted in Table CI this listener was very stable in his scoring, and the estimated error variance is thus as low as 0.28. The MS for loudspeakers (49.64) represents an estimate of the variance due to differences between the loudspeakers (the differences between the loudspeaker means in the bottom margin of Table CI) plus an estimate of the error variance. Consequently, the bigger the loudspeaker MS is relative to the within-cell MS (which represents only error variance), the more probable it is that there are “true” (non-random) differences between the loudspeakers (as scored by this listener). This is formally tested by the *F* test, which means here that the loudspeaker MS is divided by the within-cell MS, resulting in an *F* ratio of 177.29 ($= 49.64/0.28$). This value is compared with the so-called “critical” *F* value for the respective degrees of freedom (namely, $df = 3$ for loudspeakers and $df = 40$ for within-cell variation) as given in tables of the *F* distribution, which appear in most textbooks on statistics. Using a 0.01 (1%) significance level, the critical *F* value for these values of df is found to be 4.31. The obtained *F* value (177.29) is far beyond the critical value and thus significant at the 0.01 (1%) level. In the right-hand column of Table CIV this is denoted by the symbol $p < 0.01$, which means that the probability (α) of getting the observed differences between the loudspeakers solely by chance is less than 0.01. Consequently the differences may be regarded as “true”.

An analogous reasoning may be performed as regards the programmes and the interaction (loudspeakers × programmes). As seen in Table CIV the corresponding *F* tests are likewise significant at the 0.01 level.

As used here the *F* test functions as an “overall” test. For example, a significant *F* ratio as regards loudspeakers indicates only that there is at least one significant difference among all possible combinations of the loudspeakers (that is, the significant difference may be that between A and B, and/or that between A and C, and/or between A and D, B and C, B and D, C and D and/or between more complex combinations as, for example, the mean of A and C versus the mean of B and D). To find exactly which difference(s) is (are) significant, it is necessary to perform tests for specific comparisons, see Clause C4. On the other hand a non-significant *F* ratio for the loudspeakers would mean that there is no significant difference at all between the loudspeakers in any combination.

The interpretation of a significant interaction (loudspeakers × programmes) can often be reasonably made by direct inspection of the data matrix. Statistical tests for this purpose are mentioned in Clause C4. It is important to study the meaning of a significant interaction, since it may give interesting information about the performance of the different loudspeakers for different types of programme material.

C3.2 AN.VAR pour tableau de résultats de groupe

Lorsqu'une AN.VAR est effectuée sur un tableau de résultats de groupe, comme celle du tableau CII, les sources de variation vont comporter également les auditeurs et des interactions diverses faisant intervenir les auditeurs.

Les résultats obtenus par AN.VAR à partir du groupe de résultats du tableau CII sont indiqués dans le tableau suivant:

TABLEAU CV

Exemple du tableau récapitulatif pour AN.VAR effectuée sur un tableau de résultats de groupe (modèle mixte, cas 2 ci-dessous)

Source de variation	Somme de carrés (SS)	Degrés de liberté (df)	Moyenne des carrés (MS)	F	p
Haut-parleurs (L)	465,75	3	155,25	80,44	< 0,01
Programmes (P)	11,94	4	2,99	0,84	—
Auditeurs (S)	105,25	3	35,08	46,16	< 0,01
$L \times P$	47,36	12	3,95	3,09	< 0,01
$L \times S$	17,34	9	1,93	2,54	= 0,01
$P \times S$	42,86	12	3,57	4,70	< 0,01
$L \times P \times S$	45,98	36	1,28	1,68	> 0,01
à l'intérieur de la case	121,33	160	0,76		
Total	857,81	239			

En calculant la valeur F , il convient de distinguer deux cas:

- 1) Les auditeurs utilisés dans l'essai constituent, eux-mêmes, la population d'auditeurs et les résultats de l'essai ne sont strictement valables que pour eux. Le modèle statistique correspondant est appelé un modèle fixe. Dans ce cas, toutes les valeurs F sont calculées en divisant la moyenne des carrés respective MS par la moyenne à l'intérieur de la case MS (pour les haut-parleurs $155,25/0,76$, pour les programmes $2,99/0,76$, pour les auditeurs $35,08/0,76$, et ainsi de suite pour tous les termes d'interaction).
- 2) Les auditeurs sont choisis d'une manière aléatoire dans une certaine population bien définie, pour laquelle on veut généraliser les résultats de l'essai d'écoute. Les quatre auditeurs de notre exemple ont été choisis d'une manière aléatoire dans une association de « fans » de haute fidélité et peuvent donc être considérés comme représentants des membres de cette association. Le modèle statistique correspondant est appelé un modèle mixte et c'est ce modèle qui est utilisé dans le cas présent. Les valeurs F sont calculées ensuite d'une manière quelque peu différente, à savoir: le rapport F pour les haut-parleurs est obtenu en divisant la moyenne des carrés des haut-parleurs MS par la moyenne des carrés (haut-parleurs $\times$ auditeurs) MS (en abrégé, $MS_L/MS_{L \times S}$), c'est-à-dire $155,25/1,93 = 80,44$. La valeur critique pour le niveau de signification 0,01 avec le degré de liberté $df = 3$ (pour les haut-parleurs) et $df = 9$ (pour haut-parleurs $\times$ auditeurs) est 6,99. La valeur observée F est donc significative au niveau 0,01.

Le rapport F pour les programmes est obtenu d'une manière analogue en tant que $MS_P/MS_{P \times S}$, c'est-à-dire $2,99/3,57$, qui est $< 1,00$, donc non significatif (la valeur critique au niveau de signification 0,01 avec $df = 4$ et 12 est 5,41).

Le rapport F pour l'interaction (haut-parleurs $\times$ programmes) est obtenu par $MS_{L \times P}/MS_{L \times P \times S}$, c'est-à-dire $3,95/1,28 = 3,09$, significatif au niveau 0,01 (valeur critique pour df (degré de liberté) = 12 et 36 est 2,73).

C3.2 ANOVA for group data matrix

When an ANOVA is performed on a group data matrix like that in Table CII, the sources of variation will also include the listeners and various interactions involving listeners.

The results from ANOVA on the group data in Table CII are given in the following table:

TABLE CV
Example of summary table for ANOVA on group data matrix
(mixed model, case 2 below)

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F	p
Loudspeakers (L)	465.75	3	155.25	80.44	< 0.01
Programmes (P)	11.94	4	2.99	0.84	—
Listeners (S)	105.25	3	35.08	46.16	< 0.01
$L \times P$	47.36	12	3.95	3.09	< 0.01
$L \times S$	17.34	9	1.93	2.54	= 0.01
$P \times S$	42.86	12	3.57	4.70	< 0.01
$L \times P \times S$	45.98	36	1.28	1.68	> 0.01
Within-cell	121.33	160	0.76		
Total	857.81	239			

When computing the F values two cases should be distinguished:

- 1) The listeners used in the test constitute themselves the listener population, and the results from the test are strictly valid only for these listeners. The corresponding statistical model is called a fixed model. In this case all F values are computed by dividing the respective MS by the within-cell MS (for loudspeakers $155.25/0.76$, for programmes $2.99/0.76$, for listeners $35.08/0.76$, and so on for all interaction terms).
- 2) The listeners are randomly chosen from a certain defined population of listeners, to which one wants to generalize the results from the listening test. The four listeners in our example were randomly chosen from a society for high fidelity fans and could thus be considered as representative for the members of this society. The corresponding statistical model is called a mixed model, and it is this model that is used in the present case. The F values are then computed somewhat differently in the following way: The F ratio for loudspeakers is obtained by dividing the loudspeaker MS by the (loudspeakers $\times$ listeners) MS (abbreviated $MS_L/MS_{L \times S}$), that is, $155.25/1.93 = 80.44$. The critical value for 0.01 significance level with $df = 3$ (for loudspeakers) and $df = 9$ (for loudspeakers $\times$ listeners) is 6.99. The observed F value is thus significant at the 0.01 level.

The F ratio for programmes is analogously obtained as $MS_P/MS_{P \times S}$, that is $2.99/3.57$, which is < 1.00 and thus not significant (critical value at 0.01 significance level with $df = 4$ and 12 is 5.41).

The F ratio for the interaction (loudspeakers $\times$ programmes) is obtained by $MS_{L \times P}/MS_{L \times P \times S}$, that is, $3.95/1.28 = 3.09$, significant at 0.01 level (critical value for $df = 12$ and 36 is 2.73).

Les rapports F pour les auditeurs et pour toute interaction comportant des auditeurs ($L \times S$, $P \times S$ et $L \times P \times S$) sont obtenus en divisant les moyennes des carrés MS respectives par les moyennes des carrés à l'intérieur d'une case (par exemple, pour les auditeurs MS_S/MS_a à l'intérieur d'une case, $35,08/0,76 = 46,16$; pour l'interaction $L \times S$, $MS_{L \times S}/MS_a$ à l'intérieur d'une case, $1,93/0,76 = 2,54$, et ainsi de suite). Comme indiqué dans la colonne p du tableau CIV, tous ces rapports F excepté celui pour l'interaction triple $L \times P \times S$ sont significatifs au niveau 0,01 sélectionné.

Ces résultats peuvent être interprétés en bref, comme suit:

Haut-parleurs: Il existe au moins une différence significative entre les notes moyennes pour les différents haut-parleurs. Des essais pour les comparaisons spécifiques peuvent être effectués pour mettre en évidence, parmi ces différences, celles qui sont significatives (voir article C4).

Programmes: Il n'y a pas de différences significatives entre les évaluations moyennes lors des programmes différents.

Auditeurs: Il existe au moins une différence significative entre les évaluations moyennes pour les différents auditeurs (ce qui signifie qu'ils ont tendance à utiliser des parties quelque peu différentes de l'échelle 10-0; comparer, par exemple les auditeurs T et V).

Interactions: Il existe une interaction significative (haut-parleurs $\times$ programmes), c'est-à-dire que les différences entre les haut-parleurs varient d'un programme à l'autre d'une façon qui doit être étudiée plus minutieusement (éventuellement au moyen d'autres essais, voir article C4).

Les interactions significatives (haut-parleurs $\times$ auditeurs) et (programmes $\times$ auditeurs) signifient que les différences entre les différents haut-parleurs et les différents programmes varient en quelque sorte en fonction des auditeurs d'une manière qui demande une étude approfondie. Toutefois, il n'existe pas (heureusement) une triple interaction (haut-parleurs $\times$ programmes $\times$ auditeurs).

D'autres observations sur l'utilisation d'un modèle fixe ou d'un modèle mixte et les possibilités associées de généralisation sont indiquées dans le rapport technique. On y trouve également des observations sur les essais d'écoute, dans lesquels chaque auditeur ne fait qu'une évaluation pour chaque combinaison (haut-parleurs $\times$ programme).

C4. Essais pour les comparaisons spécifiques

Si le rapport F pour les haut-parleurs est significatif, il existe au moins une différence significative entre les notes moyennes pour les différents haut-parleurs (ou pour des combinaisons plus complexes de haut-parleurs). Des essais pour les comparaisons spécifiques sont réalisés pour mettre exactement en évidence, parmi ces différences, celles qui sont significatives. Malheureusement, il faut distinguer plusieurs situations différentes quand on effectue de tels essais. Quatre situations sont décrites en bref ci-dessous: comparaisons indépendantes planifiées, comparaisons dépendantes planifiées, comparaisons non planifiées et comparaisons spécifiques dans le cadre de programmes isolés. La discussion se borne, en général, à des comparaisons entre deux haut-parleurs à la fois (A par rapport à B, A par rapport à C, etc.). En outre, seul le cas d'analyse d'un groupe de résultats conformément au modèle mixte (voir paragraphe C3.2) sera discuté. D'autres cas sont traités dans le rapport technique, qui donne également des informations plus détaillées sur les fondements statistiques.

C4.1 Comparaisons indépendantes planifiées

Supposons qu'il ait été prévu, déjà avant l'essai d'écoute, qu'il était particulièrement important de vérifier si le haut-parleur A était meilleur que le haut-parleur B. Il s'agit donc dans ce cas d'une comparaison planifiée qui peut être réalisée au moyen de l'essai t . Dans le cas d'un tableau de résultats de groupe comme celui du tableau CII et en utilisant le modèle mixte, cet essai est effectué de la façon suivante: